

Faculty of Education

**A Quasi-Experimental Study of the Effects of a Reciprocal
Peer Coaching Strategy on Physiotherapy Students' Clinical
Problem Solving Skills**

Richard Kaban Ladyshevsky

**This thesis is presented as part of the requirements for
the award of the degree of Doctor of Philosophy of the
Curtin University of Technology**

March, 2000

Abstract

This research was carried out to further the theoretical and practical understanding of peer assisted learning in undergraduate physiotherapy clinical education. A quasi-experimental study, with both a control group and an experimental group, was developed to examine the effects of peer assisted learning on the cognitive, psychomotor and affective components of clinical competence.

On a more specific level, the main aim of the study was to investigate two models of clinical learning used in physiotherapy: the individualistic and reciprocal peer coaching models. Differences in history taking, physical examination and clinical reasoning were studied by having individual students, and pairs of students, complete an assessment of a simulated patient followed by a clinical reasoning test. A conceptual framework illustrates the main theoretical implications of this study. This framework incorporated current knowledge about clinical reasoning in medicine and the allied health sciences, in particular, reasoning from a novice practitioner's perspective. The conceptual framework, however, was expanded to demonstrate how peer assisted learning can influence the knowledge, cognitive and metacognitive aspects of clinical reasoning and performance.

The main findings of the study were that reciprocal peer coaching led to statistically significant improvements in performance and clinical reasoning in the experimental group. In most cases, positive effect size differences appeared in favour of the experimental group. Most noteworthy differences were in the areas of physical examination and overall thoroughness of the patient encounter. Subjects in the experimental group also outperformed subjects in the control group on the overall clinical reasoning test to a statistically significant degree. A similar outcome was also noted in terms of patient management, that is, the planning and development of treatment interventions. Low achieving students in the reciprocal peer coaching group outperformed low achieving students who worked independently. Qualitative differences in the learning atmosphere of both models were also reported by subjects. While the actual clinical reasoning process did not appear to differ across both groups, practical differences in the actual learning experience did appear. Students in

the experimental group reported being less anxious than students in the control group. The reciprocal peer coaching process was also an effective means of creating a supportive learning environment. These results are consistent with the major conclusions in the literature. Specifically, that peer assisted learning methods can increase achievement in learners (Goldschmid & Goldschmid, 1976; Johnson et al., 1998; Milson & Laatsch, 1996; Riggio et al., 1991; Riggio et al., 1994; Topping, 1996).

Acknowledgments

I wish to thank the following people most sincerely:

- my supervisor, *Associate Professor Robert Baker*, for his wisdom, advice, support and guidance *throughout the course of this study*;
- *my associate supervisors, Dr. Larry Nelson*, for his statistical advice, *Mr. Mark Jones* for his knowledge of clinical reasoning, and *Dr. Rosemary Coates*, for being there;
- the Chair of my thesis committee, *Dr. Graham Dellar*;
- *Ms. Jenny Lalor*, for her support in helping me set up and run the statistical data through SPSS;
- the *Library Staff* at Curtin University of Technology for their ongoing support throughout this study;
- *Mr. Domenic Arcorace*, for his award winning performance as my simulated patient;
- *Mr. Jeff Boyle, Mr. Andrew Walton, Ms. Nicky Fortescue, Brigitte van der Heide, and Mr. Anthony Princi* for their expert involvement in the preparation of this study;
- The numerous *fourth year physiotherapy students and post-graduate physiotherapy students* who assisted me with my pilot study and testing;
- The *academic and administrative staff at the Faculty of Education* who made my study experience enjoyable;
- *my friends*, who have had to endure the ongoing drama and saga of my studies, in particular, *Dr. Verena Marshall*, who continues to be a source of inspiration in my life, and *Molly Verrier*, who opened the door;
- the *third year physiotherapy students* who participated in this study. Their involvement was invaluable and made this thesis possible; and
- my partner, *Ron*, for his support, friendship, love and ongoing encouragement.

Table of Contents

Abstract	i
Acknowledgments	iii
Table of Contents	iv
List of Tables	x
List of Figures	xiii
Chapter 1: Introduction and Overview	1
1.1 Introduction	1
1.2 Key Objectives and Research Questions	3
1.3 Scope and Limitations	4
1.4 Research Method	5
1.5 Overview of Thesis	8
Chapter 2: Literature Review – Peer Assisted Learning	11
2.1 Peer Assisted Learning	11
2.1.1 Definition	11
2.1.2 Cooperative Learning	13
2.1.3 Peer Tutoring	18
2.1.4 Peer Collaboration	20
2.2 Peer Coaching and Professional Development	22
2.2.1 Matching Students for Reciprocal Peer Collaboration	24
2.3 Peer Assisted Learning: Theoretical Expositions of its Efficacy	26
2.3.1 Social Interdependence Theory	26
2.3.2 Cognitive-Developmental Theory	26
2.3.3 Behavioural Learning Theory	29
2.4 Peer Assisted Learning In Higher Education	30
2.4.1 Peer Assisted Learning in the Health Sciences	34
2.4.2 The Development of Generic Skills using Peer Assisted Learning	36
Chapter 3: Literature Review – Simulated Patients	38
3.1 Simulated Patients: A Definition	38
3.2 Educational Applications and Advantages of Simulated Patients	39
3.3 Simulated Patients and Evaluation of Clinical Competency	42
3.3.1 Fidelity (Content Validity) of Simulated Patients	42
3.3.2 Accuracy (reliability) of Simulated Patients	46
3.3.3 Advantages of Simulated Patients in Evaluation and Research	48
3.3.4 Training Procedures for the Simulated Patient	51
3.4 The Use of Simulated Patients in Research and Evaluation	53

3.5 The Objective Structured Clinical Examination (OSCE)	54
3.5.1 Competence and Performance Constructs	56
3.5.2 Standard Setting in Clinical Skills Evaluation.....	59
3.5.3 First Visit Bias	61
3.5.4 Station Duration.....	62
3.5.5 Using Two or More Simulated Patients to Portray a Single Case.....	66
3.5.6 Gender Bias in Evaluations using Simulated Patients	67
3.5.7 Sequential Testing and Examination Security	69
3.5.8 Case Specificity and the Generalisability of Findings in Tests of Clinical Competence/Performance.....	73
3.5.9 Student Anxiety in Performance Based Testing	77
3.6 Methods and Principles for the Evaluation of Clinical Competence in Objectively Structured Clinical Examinations	80
3.6.1 Station Development.....	80
3.6.2 Subjective versus objective measurement instruments.....	81
3.6.3 Rater Reliability in Performance Based Examinations.....	85
3.6.4 Methods (instruments) for the evaluation of history-taking, physical examination and communication skills.....	91
3.6.5 Checklists and Rating Scales	92
3.6.6 Methods (instruments) for the evaluation of post-encounter clinical reasoning	96
3.7 Measures of Validity for Simulated Patient Examinations	101
3.7.1 Past Experience	101
3.7.2 Evaluating Validity in Performance Based Assessment.....	101
3.7.3 Content Validity.....	102
3.7.4 Construct Validity	103
3.7.5 Criterion Validity	106
3.7.6 Internal Consistency.....	109
Chapter 4: Literature Review – Clinical Reasoning	111
4.1 Clinical Reasoning	111
4.1.1 Definitions.....	111
4.2 Clinical Reasoning in Medicine	113
4.2.1 The Hypothetico-Deductive Reasoning Model.....	113
4.2.2 The Representation of Knowledge in the Memory of Clinicians	120
4.2.3 Studies of Novice and Expert Problem Solving in Chess and Physics.....	122
4.2.4 Pattern Recognition in Medicine	124
4.2.5 Forward and Backward Reasoning	127
4.2.6 Organisation of Knowledge as Prototypes.....	129
4.2.7 Illness Scripts	131
4.2.8 An Emerging Theory of Expertise in Medicine.....	133

4.2.9 The Influence of Biomedical and Clinical Knowledge.....	135
4.2.10 Errors in Clinical Reasoning.....	137
4.3 Clinical Reasoning in the Allied Health Professions.....	140
4.4 Clinical Reasoning in Occupational Therapy	140
4.5 Clinical Reasoning in Nursing	144
4.6 Clinical Reasoning in Physiotherapy	148
4.6.1 Diagnosis and Physiotherapy	148
4.6.2 Hypothetico-Deductive Reasoning in Physiotherapy.....	149
4.6.3 Phenomenological Approaches to Reasoning in Physiotherapy	152
4.6.4 Pattern Recognition in Physiotherapy.....	154
4.6.5 Clinical Reasoning of Physiotherapy Students	156
4.7 Clinical Reasoning and the Novice Practitioner	158
4.7.1 The Novice: A definition	158
4.7.2 The Novice Practitioner: Clinical Reasoning Characteristics.....	159
4.8 Educational Implications of the Clinical Reasoning Literature	162
4.8.1 The Patient Evaluation Process.....	163
4.8.2 Experiential Learning: A Constructivist Perspective For Teaching and Learning in Physiotherapy	164
4.8.3 Peer Discussion.....	167
4.9 Metacognition and the Management of Knowledge	169
4.9.1 Definitions of Metacognition.....	169
4.9.2 The Influence of Metacognition on Learning	171
4.10 Qualitative Research Methods to Study Clinical Reasoning	174
4.10.1 Using Verbal Reports to Study Clinical Reasoning.....	175
4.10.2 Theoretical Foundations for the Use of Verbal Reports as Data	175
4.10.3 Criticisms and Limitations of Verbal Report Data	178
4.10.4 The Influence of Verbalisation on Task Performance	179
4.10.5 Stimulated Recall and Studies of Clinical Reasoning.....	180
4.11 Methods for the Analysis of Qualitative Data or Verbal Records	183
4.11.1 Multiple Cases	183
4.11.2 Coding.....	183
4.11.3 Coding Accuracy and Reliability	185
4.12 Characteristics and Limitations of Qualitative Data Analysis	186
Chapter 5: Conceptual Framework.....	188

Chapter 6: Methods	190
6.1 Research Design	190
6.1.1 The Simulated Patient	190
6.1.2 Case Selection for this Study	190
6.1.3 Development of the SP Case.....	191
6.1.4 Developing the Measurement Instruments for the SP Case.....	192
6.2 Face and Content Validity of the SP Case and Test Materials.....	194
6.2.1 Review of the History and Physical Examination Checklists	195
6.2.2 Review of the Arizona Clinical Interviewing Rating Scale	196
6.2.3 Post-Encounter Questionnaire Review	196
6.2.4 Review of the Simulated Patient Training Notes.....	197
6.3 Simulated Patient Recruitment and Training	197
6.3.1 Recruitment and Selection Issues	197
6.3.2 Simulated Patient Training	197
6.3.3 Practice Sessions	198
6.4 The Pilot Test	198
6.4.1 Subjects	199
6.4.2 The Simulated Patient Encounter.....	199
6.4.3 Post-Encounter Activities	200
6.4.4 Determining Station Duration.....	201
6.4.5 Accuracy, Inter-Rater & Intra-Rater Reliability of the Simulated Patient	201
6.4.6 Construct Validity of the Measurement Instruments	204
6.4.7 Criterion Validity - concurrent.....	206
6.4.8 Review of the Measurement Instruments.....	207
6.4.9 Qualitative Aspects of Clinical Reasoning	212
6.5 Main Study	213
6.5.1 The Context.....	213
6.5.2 Sampling and Allocation to IND and RPC Groups	214
6.5.3 Grapevine Effect	221
6.5.4 Preparation of Students for Reciprocal Peer Coaching.....	222
6.5.5 The Simulated Patient Encounter.....	223
6.5.6 Post-Encounter Activities	224
6.6 Data Analysis	225
6.6.1 Research Objective 1	227
6.6.2 Research Objective 2	229
6.6.3 Research Objective 3	230
6.7 Qualitative Methods.....	231
6.7.1 Stimulated Recall	231
6.7.2 Stimulated Recall: Sampling Methods.....	231
6.7.3 Stimulated Recall: Specific Methods.....	232

Chapter 7: Results	237
7.1 Research Objective 1:	237
7.2 Research Questions 1 – 3:	237
7.3 Research Question 4:	241
7.3.1 Time to Complete the Patient Encounter	241
7.3.2 Thoroughness of the IND and RPC Groups During the SP Encounter	243
7.3.3 Efficiency of the IND and RPC Groups During the SP Encounter	246
7.4 Research Objective 2:	247
7.5 Research Question 5:	247
7.5.1 Performance on the Post-Encounter Clinical Reasoning Questionnaire for the IND and RPC Groups.....	247
7.5.2 Differences in “Low Student” Performance on the Post- Encounter Clinical Reasoning Questionnaire Across the IND and RPC Groups.....	251
7.5.3 Differences in “High Student” Performance on the Post- Encounter Clinical Reasoning Questionnaire Across the IND and RPC Groups.....	255
7.6 Research Question 6:	259
7.6.1 Performance on All Measures Within the RPC Group.....	259
7.6.2 Performance on All Measures Within the IND Group.....	268
7.7 The Relationship Between Performance on the Patient Encounter and Post-Encounter Reasoning	271
7.8 Research Objective 3	273
7.9 Research Question 7:	273
7.9.1 Anxiety and Confidence Scores of the Subjects	273
7.10 Features of the Simulated Patient Case	275
7.10.1 Accuracy of the Simulated Patient (SP).....	275
7.10.2 Inter-rater Reliability of the Simulated Patient	275
7.10.3 Grapevine Effects.....	276
7.11 Research Question 8:	278
7.11.1 Clinical Reasoning	279
7.11.2 The Learning Experience	289
7.11.3 The Reciprocal Peer Coaching Experience.....	291
7.11.4 The Reciprocal Peer Coaching Experience - Student Evaluation	296
Chapter 8: Discussion	301
8.1 Summary of Findings	301
8.1.1 Research Objective 1:	301
8.1.2 Research Question 1:	301

8.1.3 Research Question 2:	302
8.1.4 Research Question 3:	302
8.1.5 Research Question 4:	303
8.1.6 Research Objective 2:	305
8.1.7 Research Question 5:	305
8.1.8 Research Question 6:	310
8.1.9 Research Objective 3:	312
8.1.10 Research Question 7:	312
8.2 Limitations	315
8.2.1 The Simulated Patient and the Simulation	315
8.2.2 First Visit Bias	318
8.2.3 Gender Bias	319
8.2.4 Grapevine Effect	319
8.2.5 The 15 Minute Discussion Period	320
8.2.6 Retrospective Recall	321
8.2.7 Singular Case	323
Chapter 9: Conclusions	324
9.1 Reciprocal Peer Coaching - Relationship to the Literature	324
9.1.1 Reward Systems to Promote Peer Coaching	325
9.1.2 The Value of Discussion and Peer Interaction on Metacognition	326
9.1.3 Learning From Peers	329
9.1.4 Training	332
9.1.5 Competition	334
9.2 The Teaching of Clinical Reasoning	334
9.3 Future Research Implications	339
Appendix 1 : History Checklist	380
Appendix 2 : Physical Examination Checklist	383
Appendix 3 : Arizona Clinical Interview Rating Scale	387
Appendix 4 : Post-Encounter Probe	388
Appendix 5 : Simulated Patient Training Notes	394
Appendix 6 : Pre-Encounter Questionnaire	404
Appendix 7 : Post-Encounter Questionnaire	405
Appendix 8 : Demographic Questionnaire	406
Appendix 9 : Instructions for Recall Session	407
Appendix 10 : Instructions for Warm-Up Sessions	408
Appendix 11 : Data Screening	410

List of Tables

Table 3.1: Educational Uses of Simulated Patients	40
Table 3.2: Summary of Station Duration in SP Evaluations	64
Table 4.1: Categories of Types of Thinking in Clinical Problem Solving.....	117
Table 4.2: Learned Capabilities or Outcomes	121
Table 4.3: Clinical Reasoning Errors.....	139
Table 4.4: The Nature of Expertise.....	162
Table 4.5: Approaches to Learning.....	173
Table 4.6: Studies Employing Retrospective (R) or Concurrent (C) Verbalisations for the Study of Clinical Reasoning	182
Table 6.1: Panel Members.....	195
Table 6.2: Fidelity of the Simulated Patient	200
Table 6.3: Inter-rater and Intra-rater Reliability of the Simulated Patient.....	203
Table 6.4: Independent Samples t-test for Mean Pre-Anxiety/Confidence	205
Table 6.5: Independent Samples t-test for Mean Scores on the Patient Encounter Checklist and Post-encounter Questionnaire	205
Table 6.6: Item Analysis for the History Checklist and Physical Examination Checklist	208
Table 6.7: Item Analysis for the Post-Encounter Questionnaire: Key Feature Questions 4 - 8.....	211
Table 6.8: Demographics of IND and RPC Group	216
Table 6.9: Background Shoulder Pathology Experience of IND and RPC groups (Observation).....	219
Table 6.10: Background Shoulder Pathology Experience of IND and RPC groups (Assessment).....	220
Table 6.11: Background Shoulder Pathology Experience of IND and RPC Groups (Treatment).....	221
Table 6.12: Mean Orthopaedic Scores of Subjects Week by Week	222
Table 6.13: Clinical Reasoning Codes.....	234
Table 6.14: Learning Experience Codes: Generic	235
Table 6.15: Reciprocal Peer Coaching Codes	235
Table 7.1: Independent Samples t-tests on History, Physical Examination and Communication Scores between the IND and RPC Groups	238
Table 7.2: Range of History, Physical Examination and Communication Raw Scores for the IND and RPC Groups	238
Table 7.3: Independent Samples t-test for Time to Complete the Patient Encounter: IND and RPC Groups	243
Table 7.4: Range of Times to Complete the Patient Encounter: IND and RPC Groups.....	243

Table 7.5: Independent Samples t-test for Mean Thoroughness of History and Physical Examination: IND and RPC Groups.....	245
Table 7.6: Range of Thoroughness Scores for History and Physical Examination: IND and RPC Groups.....	246
Table 7.7: Independent Samples t-test for Mean Efficiency Scores for History and Physical Examination: IND and RPC Groups.....	246
Table 7.8: Range of Efficiency Scores for History and Physical Examination: IND and RPC Groups.....	246
Table 7.9: Independent Samples t-test for Outcome Scores for the Post-Encounter Questionnaire: IND and RPC Groups.....	251
Table 7.10: Range of Outcome Scores for the Post-Encounter Questionnaire: IND and RPC Groups.....	251
Table 7.11: Independent Samples t-tests for PEQ Outcome Scores: IND and RPC Groups – Low Students.....	254
Table 7.12: Range of PEQ Outcome Scores: IND and RPC Groups – Low Students.....	255
Table 7.13: Independent Samples t-test for PEQ Outcome Scores: IND and RPC Groups – High Students.....	258
Table 7.14: Range of Outcome Scores for the Post-Encounter Questionnaire: IND and RPC Groups - High Students.....	259
Table 7.15: One Way ANOVA for Outcomes Scores for the SP Encounter: RPC Group.....	267
Table 7.16: One Way ANOVA for Outcomes Scores for Post-Encounter Questionnaire: RPC Group.....	268
Table 7.17: Independent Samples t-test for Outcome Scores on All Measures for High and Low Students: IND Group.....	270
Table 7.18: Correlations between the History and Physical Examination Checklists and the Post-Encounter Questionnaire.....	271
Table 7.19: Paired Sample t-tests for Anxiety and Confidence Measures for the IND Group.....	274
Table 7.20: Paired Sample t-tests for Anxiety and Confidence Measures for the RPC Group.....	274
Table 7.21: Inter-rater Reliability of the Simulated Patient.....	276
Table 7.22: Group Composition Structure Across the 4 Weeks.....	276
Table 7.23: History and Physical Examination Checklist Scores Across the Duration of the Study.....	277
Table 7.24: Post-Encounter Questionnaire Score Across the Duration of the Study.....	277
Table 7.25: Mean Scores and Range of Scores for Clinical Reasoning Coding Categories by IND and RPC Group.....	284
Table 7.26: Absolute Number of Items within a Coding Category: IND and RPC Group.....	285
Table 7.27: Actual Number of Items Within the Source of Symptom Category: IND and RPC Groups.....	286

Table 7.28: Actual Number of Items Within the Mechanism of Signs and Symptoms Category: IND & RPC Groups	286
Table 7.29: Actual Number of Items Within the Contributing Factors Category: IND and RPC Groups	287
Table 7.30: Actual Number of Items Within the Naming a Hypothesis Category: IND and RPC Groups	287
Table 7.31: Appearance of Hypotheses in Relation to the Overall Encounter: IND Group	288
Table 7.32: Appearance of Hypotheses in Relation to the Overall Encounter: RPC Group	289
Table 7.33: Learning Experience Codes: IND and RPC Groups	291
Table 7.34: Reciprocal Peer Coaching Experience Codes: IND and RPC Groups.....	295

List of Figures

Figure 1: Flow chart of the research design used in the study	6
Figure 2.1: The Process of Controversy	28
Figure 4.1: The Clinical Reasoning Process.....	116
Figure 4.2: The Clinical Reasoning Process.....	118
Figure 4.3: Analysis of the Hypothetico-Deductive Reasoning Model	119
Figure 4.4: Analysis of the Pattern Recognition Model (Forward/Inductive Reasoning) Source: Higgs & Jones (1995)	128
Figure 5.1: Conceptual Framework	189
Figure 6.1: Reciprocal Peer Coaching Questions.....	213
Figure 6.2: Distribution of Third Year Physiotherapy Students OS 252/351 Grades (n=62).....	216
Figure 6.3: Background Shoulder Pathology Experience of IND (n=20) and RPC (n=42) groups (Observation)	218
Figure 6.4: Background Shoulder Pathology Experience of IND (n=20) and RPC (n=42) groups (Assessment).....	219
Figure 6.5: Background Shoulder Pathology Experience of IND (n=20) and RPC (n=42) groups (Treatment).....	220
Figure 7.1: Boxplots for History Checklist (HCList): IND and RPC Groups.....	239
Figure 7.2: Boxplots for Physical Examination Checklist (PECList): IND and RPC Groups.....	240
Figure 7.3: Boxplots for the Arizona Clinical Interview Rating Scale (ACIRS): IND and RPC Groups.....	240
Figure 7.4: History, Physical Examination and Communication Scores for the IND and RPC Groups.....	241
Figure 7.5: Boxplots for Time (in minutes) to Complete the History (TimeHx) for the IND and RPC Groups	242
Figure 7.6: Boxplots for Time (minutes) to Complete the Physical Examination (TimePEX) for the IND and RPC Groups.....	242
Figure 7.7: Mean Time (in minutes) to Complete the SP Encounter: IND and RPC Groups.....	242
Figure 7.8: Boxplots for Thoroughness of Patient History (ThoroHx): IND and RPC Groups.....	244
Figure 7.9: Boxplots for Thoroughness of Physical Examination (ThoroPex): IND and RPC Groups.....	244
Figure 7.10: Mean Thoroughness of History and Physical Examination: IND and RPC Groups.....	245
Figure 7.11: Boxplots for PEQ Diagnosis Score (PEQDx): IND and RPC Groups.....	248
Figure 7.12: Boxplots for PEQ Management Score (PEQMgmt): IND and RPC Groups.....	248

Figure 7.13: Boxplots for PEQ History Key Features Score (PEQHxKF): IND and RPC Groups.....	249
Figure 7.14: Boxplots for PEQ Physical Examination Key Features (PEQPExKF): IND and RPC Groups	249
Figure 7.15: Boxplots for PEQ Total Score (PEQTotal): IND and RPC Groups	249
Figure 7.16: Performance Scores on the PEQ for the IND and RPC Groups	250
Figure 7.17: Boxplots for Post-Encounter Clinical Reasoning Test: Diagnosis Scores (PEQDx) of Low Students in the IND and RPC Groups.....	252
Figure 7.18: Boxplots for Post-Encounter Clinical Reasoning Test: Management Scores (PEQMgmt) of Low Students in the IND and RPC Groups.....	252
Figure 7.19: Boxplots for Post-Encounter Clinical Reasoning Test: History Key Feature Scores (PEQHxKF) of Low Students in the IND and RPC Groups.....	253
Figure 7.20: Boxplots for Post-Encounter Clinical Reasoning Test: Physical Examination Key Feature Scores (PEQPExKF) of Low Students in the IND and RPC Groups	253
Figure 7.21: Boxplots for Post-Encounter Clinical Reasoning Test: Total Score (PEQTot) of Low Students in the IND and RPC Groups	253
Figure 7.22: Performance Scores on the PEQ for Low Students: IND and RPC Groups.....	254
Figure 7.23: Boxplots for Post-Encounter Clinical Reasoning Test: Diagnosis Scores (PEQDx) of High Students in the IND and RPC Groups	255
Figure 7.24: Boxplots for Post-Encounter Clinical Reasoning Test: Management Scores (PEQMgmt) of High Students in the IND and RPC Groups.....	256
Figure 7.25: Boxplots for Post-Encounter Clinical Reasoning Test: History Key Feature Scores (PEQHxKF) of High Students in the IND and RPC Groups.....	256
Figure 7.26: Boxplots for Post-Encounter Clinical Reasoning Test: Physical Exam. Key Feature Scores (PEQPExKF) of High Students in the IND and RPC Groups.....	257
Figure 7.27: Boxplots for Post-Encounter Clinical Reasoning Test: Total Score (PEQTot) of High Students in the IND and RPC Groups.....	257
Figure 7.28: Performance Scores on the PEQ for High Students: IND and RPC Groups.....	258
Figure 7.29: Boxplots for the High:High Group - Time and Checklist Measures for the SP encounter.....	261
Figure 7.30: Boxplots for High:Low Group - Time and Checklist Measures for the SP Encounter	262
Figure 7.31: Boxplots for Low:Low Group - Time and Checklist Measures for the SP Encounter	263
Figure 7.32: Boxplots for High:High Group - PEQ Outcomes.....	264
Figure 7.33: Boxplots for High:Low Group - PEQ Outcomes.....	265
Figure 7.34: Boxplots for Low:Low Group - PEQ Outcomes	265

Figure 7.35: Time (in minutes) to Complete the SP Encounter for High:High, High:Low and Low:Low RPC Groups	266
Figure 7.36: Checklist Scores for the SP Encounter for the High:High, High:Low and Low:Low RPC Groups	266
Figure 7.37: Performance on the PEQ for the High:High, High:Low and Low:Low RPC Groups.....	267
Figure 7.38: Time (in minutes) to Complete the SP Encounter for High and Low Students: IND Group	269
Figure 7.39: Checklist Scores for the SP Encounter for High and Low Students: IND Group.....	269
Figure 7.40: Outcome Scores on the PEQ for High and Low Students: IND Group.....	270
Figure 7.41: Scatterplot of Relationship between History and Physical Examination Checklist Score to Post-Encounter Questionnaire: Management Question Score	272
Figure 7.42: Scatterplot of Relationship between History and Physical Examination Checklist Score to Overall Post-Encounter Questionnaire Score	272
Figure 7.43: Pre- and Post- Anxiety Scores for Both Groups	275
Figure 7.44: Pre- and Post- Confidence Scores for Both Groups.....	275
Figure 7.45: Linear Representation of Weekly Scores	278
Figure 7.46: Source of Symptoms: Examples from Subjects.....	279
Figure 7.47: Mechanism of Signs and Symptoms: Examples from Subjects.....	280
Figure 7.48: Contributing Factors: Examples from Subjects	280
Figure 7.49: Naming a Hypothesis: Examples from Subjects.....	280
Figure 7.50: Precautions and Contraindications: Examples from Subjects.....	281
Figure 7.51: Minimising Clinical Reasoning Errors: Examples from Subjects	281
Figure 7.52: Management of the Client: Examples from Subjects	282
Figure 7.53: Prognosis: Examples from Subjects	282
Figure 7.54: Psychosocial Inquiry: Examples from Subjects.....	283
Figure 7.55: Knowledge Gap: Examples from Subjects	289
Figure 7.56: Acknowledging Anxiety: Examples from Subjects.....	290
Figure 7.57: Lacking Confidence: Examples from Subjects.....	290
Figure 7.58: Simulation Noted: Examples from Subjects	290
Figure 7.59: Determining Roles: Examples from Subjects	292
Figure 7.60: Affirming Action: Examples from Subjects	292
Figure 7.61: Critiquing Action: Examples from Subjects	293
Figure 7.62: Seeking Clarification: Examples from Subjects	293
Figure 7.63: Intercepting a Course of Action Positively: Examples from Subjects.....	294
Figure 7.64: Support: Examples from Subjects	294
Figure 7.65: Competition: Examples from Subjects.....	295

Chapter 1: Introduction and Overview

1.1 Introduction

Kaufman, Portney, & Jette (1997) have stated that physiotherapy educators need to develop systems which foster the development of therapists who are able to use a broad body of knowledge to solve complex clinical problems. The development of competent practitioners, therefore, is a major goal of academic and clinical education programs. To achieve this competency, students in health science programs spend a significant amount of time engaged in clinical activities as part of their studies. Many programs in Australia, for example, spend up to one third of their curricula in clinical education (Ferguson & Edwards, 1999). The structure of these clinical education programs also differ across States, within universities and across professional disciplines (Ferguson & Edwards, 1999).

At Curtin University of Technology, undergraduate physiotherapy students spend approximately 950 hours in clinical settings. Much of this clinical education is overseen by a registered physiotherapist who serves as an expert role model and supervisor. This supervisor's role is to monitor the performance of the student and to provide feedback. This model of supervision puts a large onus on the supervisor to provide mentorship and developmental support to the student. It is also a very resource intensive training model requiring one licensed physiotherapist per student. In many ways, this clinical education model mirrors apprenticeship. Apprenticeship models, however, may not be the best way to educate a health care practitioner (Gerace & Sibilano, 1984; Higgs & Hunt, 1999; Tinning & Fitzclarence, 1982). These educators argue that in the apprenticeship model, students are required to plan their learning experiences under the direction of their supervisor. Opportunities for planning with peers is minimal and often leads to a reliance on others with higher role status. Further, apprenticeship models appear to force students to rely on experts who transmit knowledge and direct learning (Boud & Edwards, 1999; Carpenter, 1996; Damon & Phelps, 1989; Lynch, 1984; Williams, 1995). Students, as a result, are often denied the opportunity to develop their own resources, to interact effectively, or given an opportunity to deal with disagreements. The outcome, therefore, is a graduate who is technically sound but lacking in some of the generic

skills required in the workforce. This is not to say that apprenticeship styles of learning should be abandoned. There are benefits to this teaching and learning model which cannot be ignored (Perry, 1988). For example, students are required to perform independently and to make individual decisions. These are pre-requisites for autonomous professional practice. Students also receive individual tuition and have opportunities to observe and learn about expert models of practice (Higgs & Hunt, 1999; Paschal & Jensen, 1995; Perry, 1988).

There is evidence in the literature, however, which suggests that peers can learn a great deal from one another (Antil, Jenkins, & Wayne, 1998; Cohen & Sampson, 1999; Fantuzzo, 1989; Johnson, Johnson, & Smith, 1998; Lake, 1999; Topping, 1996). If this is the case, then reliance on the expert clinician could be reduced. By creating opportunities in the clinical setting which permit peer assisted learning to take place, an opportunity exists to enhance the overall professional development experience (Williams, 1995). To this end, there is no reason why peer assisted learning (PAL) cannot co-exist with more traditional forms of learning in the clinical setting (Goldenberg & Iwasiw, 1992). In fact, PAL should be encouraged as it is one method of promoting achievement in students. There is growing evidence that this learning method produces significant achievement at the tertiary level. Costello (1989), for example, states that students report learning most of their practical skills from their colleagues or peers even though they had faculty supervisors. Costello (1989) refers to this as the 'hidden curriculum' and describes it as student learning which occurs over and above the structured curriculum, or that provided by a clinical educator.

Surprisingly, the use of PAL is not new. Over twenty years ago it was suggested that peer instruction in the clinical setting was a useful method for developing management and teaching skills in students, without decreasing their exposure to direct client care (Cason, Cason, & Bartnik, 1977). With increasing limits on available resources in the health care system, the use of PAL is now being seen as a more attractive option. Increasing workplace demands on clinical instructors often means that students receive inadequate amounts of feedback on their performance. This lack of supervisory feedback typically stems from staff shortages and workplace demands which makes high quality 'one on one' supervision increasingly difficult.

This sentiment has certainly been echoed in North America (Haffner-Zavadak, Konecky-Dolnack, Polich, & Van Volkenburg, 1995) and does not appear to differ from what is happening in Australia (Ferguson & Edwards, 1999).

Despite shrinking clinical education resources and workplace pressures, many allied health programs still rely on the 'one-on-one' supervisory model (DeClute & Ladyshevsky, 1993; DeDea, 1996; Williams, 1995). This may not be an acceptable strategy given that the cost-effectiveness and quality of teaching is under greater scrutiny at a time when larger enrolments and fewer resources are available (Gandy, 1995; Hill, Gay, & Topping, 1998). In response to these pressures, new models of clinical education are being developed (Carpenter, 1996). The reciprocal peer coaching model under investigation in this thesis is one such alternative.

1.2 Key Objectives and Research Questions

Two models of learning in a clinical education setting are being investigated in this study. These two models are the individualistic (IND) and reciprocal peer coaching (RPC) learning models. The individualistic model involves a student working through a clinical problem and is one of the most common and traditional forms of clinical learning in physiotherapy (Gandy, 1999). The student's learning goals in this situation are generally unrelated and independent from other students. The RPC model involves two students working through a clinical problem collaboratively. This model of learning is less common in the clinical education setting. In RPC, the students' goals are inter-related and a successful outcome is dependent upon them being able to coach and support one another.

There are three main objectives for this study which have been used to guide this research. These objectives focus on the three domains of clinical competence: psychomotor, cognitive and affective (Bloom et al., 1956; Shepard & Jensen, 1999). The associated research questions which accompany each of these objectives follow.

Objective 1: To determine differences in undergraduate physiotherapy student performance, from the perspective of a patient encounter, across the IND and RPC learning models.

- **Research Question 1:** How does history taking skill differ in the IND and RPC Group?
- **Research Question 2:** How does physical examination skill differ in the IND and RPC Group?
- **Research Question 3:** How does interviewing skill differ in the IND and RPC Group?
- **Research Question 4:** Is there a difference in the time it takes to complete the task, thoroughness, and efficiency, across the IND and RPC Groups?

Objective 2: To determine differences in clinical reasoning and problem solving across the IND and RPC learning models.

- **Research Question 5:** Are there differences in clinical reasoning across the IND and RPC group from the perspective of diagnostic skill, ability to identify management options for the client, and skill in identifying the key features of the case?
- **Research Question 6:** Are there differences within the IND and RPC group in terms of their history, physical examination, interviewing and clinical reasoning skills?

Objective 3: To determine differences in the affective domain of clinical competence across the IND and RPC learning models.

- **Research Question 7:** How is anxiety and confidence influenced by the IND and RPC learning models?
- **Research Question 8:** What experiences do students report when engaged in an IND and RPC learning model?

Another objective of this study, albeit a very minor one, was to determine whether a full scale patient simulation could be used with confidence in studies of clinical reasoning and performance in physiotherapy.

1.3 Scope and Limitations

Controlled investigations of different clinical education learning models are lacking in the physiotherapy education literature. Hence, this study will make a significant

contribution to our understanding of the differences between individualistic and peer assisted learning when employed in the clinical setting: in particular, the effects on student performance and problem solving.

As with any research, however, there are limitations. First of all, this study only examines physiotherapy student performance and reasoning in orthopaedics. Second, only one patient case is used to investigate differences across the IND and RPC groups. This further limits the generalisability of the results. Hence, this study should be seen as expanding upon an area of investigation for researchers interested in exploring clinical reasoning and performance, in particular, the effect that different learning models have on the development of these skills.

1.4 Research Method

A quasi-experimental design was used to investigate the research questions for this study. Both an experimental and a control group were exposed to the independent variable, a simulated patient, and a series of instruments were developed to measure the influence of the IND and RPC models on the dependent variables. Throughout the study, steps were taken to ensure that the experiment was carried out in an ethical manner. The study was approved by Curtin University of Technology's Ethics Committee. Students were assured that the information obtained on all aspects of this study would be managed confidentially. Anonymity was ensured at all times.

In order to gain a picture of how this study was carried out, a flow chart of the various parts of the study is illustrated in Figure 1.1.

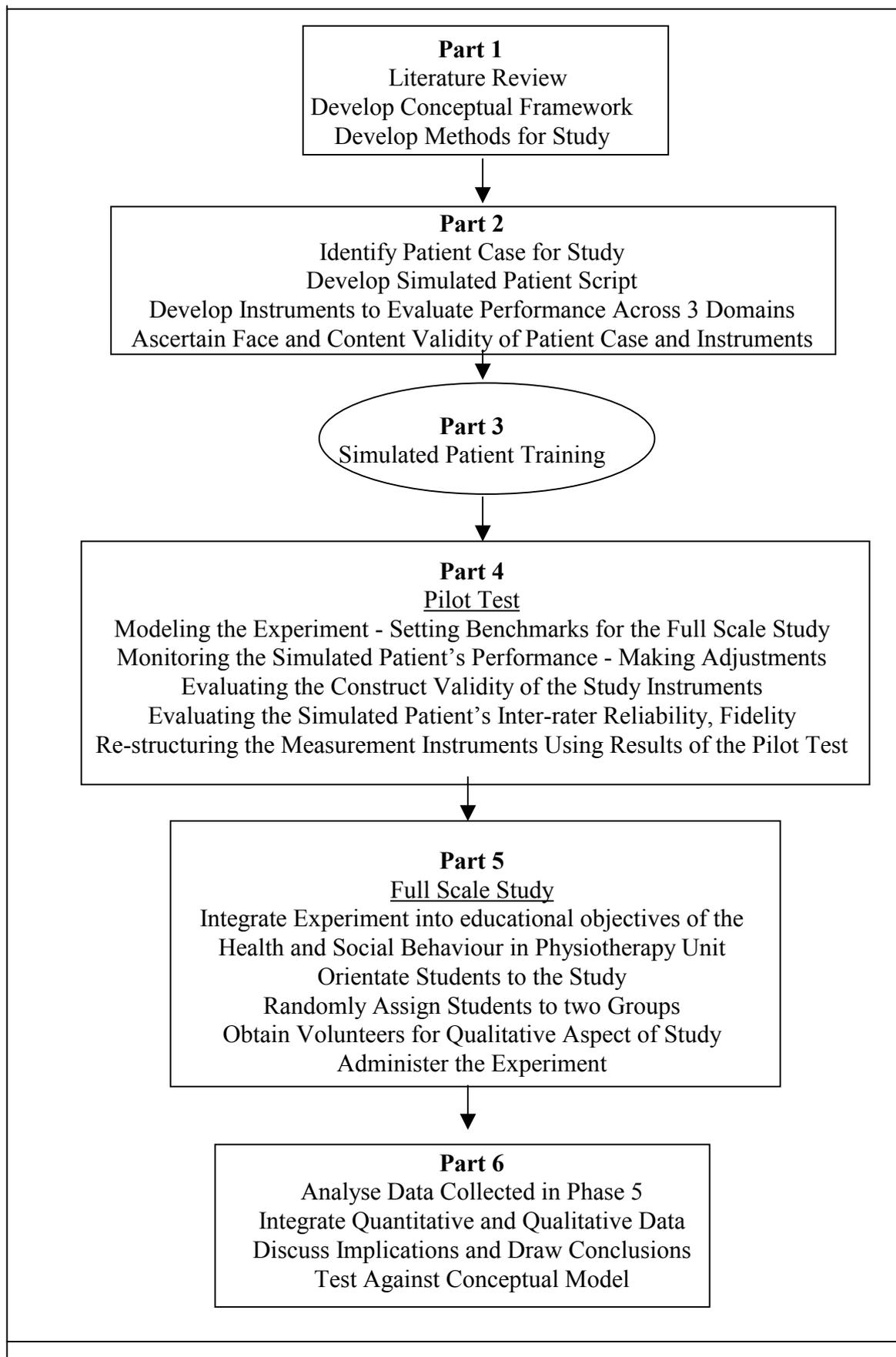


Figure 1: Flow chart of the research design used in the study

The first part of this study involved a review of the literature from education, medicine, allied health sciences and psychology. There were three foci to this review. The first focus was on the peer assisted learning literature. Because this literature traverses primary through to tertiary levels of education, the review focused on peer assisted learning in the higher education sector. The second focus of the review centered around the measurement of clinical competence and performance. Given the aims and objectives of this research, accurate, reliable and valid measures of performance were needed. The final focus of the literature review looked at clinical reasoning from the perspective of the health professions. Differences in novice and expert reasoning and the educational implications for enhancing this skill were the major foci of this review. The conceptual framework which guided this research, along with the methods to validate this framework, followed from this review of the literature.

The second part of this study focused on developing the materials and measurement instruments for the experiment. During this part of the study the simulated patient case was conceptualised and developed. The measurement instruments to measure history taking, physical examination, interviewing, and clinical reasoning skill were also created and subjected to face and content validation.

The third part of this study involved training an actor to portray the paper based simulated patient. Role plays with the investigator, senior physiotherapy students and post-graduate physiotherapists were carried out in an effort to make the simulated patient's performance realistic and valid. During these training sessions, the simulated patient was also trained to rate the candidate's performance so that his ratings were valid and reliable.

Part four of the study was a small scale version of the full study. This pilot test was carried out to set the benchmarks for the full scale study and to iron out any wrinkles associated with the methods. The pilot test also measured the construct and concurrent validity and reliability of the measurement instruments along with the performance and rating accuracy of the simulated patient.

The fifth part of this research was the full scale implementation of the experiment. The subjects that participate in this study were 62 undergraduate physiotherapy

students at the end of their third year of study at Curtin University of Technology. These students were randomly assigned to either the experimental or control group and carried out a full history and physical examination of the simulated patient. Their performance and clinical reasoning skill was measured using the specially designed instruments and processes for this study.

The final part of this study involves analysing the quantitative and qualitative data from the experiment and drawing conclusions about the outcomes.

1.5 Overview of Thesis

This thesis is organised into nine different chapters, including this first Introductory Chapter. Chapters 2 to 4 are the literature review and are substantive out of necessity to address the various perspectives of this research. The literature review covers a variety of topics. These are listed below.

1. Peer assisted learning
2. Problem based learning
3. Patient Simulation
4. The psychometrics of performance evaluation in the health sciences
5. Evaluation of cognition and thought processes in clinical diagnosis
6. Clinical reasoning
7. Teaching and learning in the health sciences.

The first part of the literature review in Chapter 2 focuses on peer assisted learning. A substantial part of this review is designed to clarify the numerous terms that are used to describe peer assisted learning. The theoretical background to peer assisted learning is also explored with connections being made to professional development.

Chapter 3 of the literature review provides a more in-depth look at the methods associated with the measurement and evaluation of clinical performance and problem solving. A comprehensive overview of simulated patient technology is also provided, in particular, the validity and reliability of the simulated patient and the suitability of using this technology for research investigations. The next part of this review focuses on methods to evaluate or measure clinical performance and reasoning. Many of the

measurement instruments or techniques that are described in this chapter are used in the context of an objectively structured clinical examination. Others are used to measure knowledge or clinical reasoning.

Given that one of the central goals of this thesis is to examine how different learning models influence clinical problem solving, Chapter 4 explores the literature on clinical reasoning. This literature is drawn from the fields of medicine, nursing, occupational therapy and physiotherapy. From here, one can then examine the literature on novice reasoning and the educational implications for developing mastery. In particular, the relationship between peer assisted learning, clinical reasoning and the development of higher order cognitive skills.

As there are qualitative methods employed in this thesis to study problem solving, the final part of Chapter 4 provides a review of the literature on the qualitative analysis of clinical reasoning. The theoretical foundations of stimulated recall, and how it can be used to study clinical reasoning are the focus of this review.

Chapter 5 provides an explanation of the conceptual framework for this study. It is followed by Chapter 6 which describes the methods that have been employed in this study. This chapter is broken down into five general sections. The first section describes the methods that were used to develop the simulation and testing instruments. The second section describes the pilot study. A pilot study was carried out to determine the reliability and validity of the simulated patient and measurement instruments. Issues associated with the simulation itself were also reviewed during the pilot testing. The third section describes the full scale study and overviews subject selection and allocation to the experimental and control groups. The fourth section describes the methods that were employed to analyse the data. The fifth section describes the methods used to manage the stimulated recall component of the study.

Chapter 7 is a summary of the results of the current study. There are two parts to this chapter. The first part represents the quantitative measurement outcomes of the study while the second part represents the qualitative measurement outcomes. Chapters 8 and 9 are the discussion and conclusion chapters of this thesis respectively. In the discussion chapter, the outcomes of this study, along with some of its limitations, are

more thoroughly discussed. The conclusion chapter reports on the educational implications of this study from a theoretical perspective and suggests future areas of research.

List of Abbreviations Used

ACIRS	Arizona Clinical Interviewing Rating Scale
CR	Clinical Reasoning
HC	History Checklist
H-D	Hypothetico-Deductive
HSBIP	Health and Social Behaviour in Physiotherapy
IND	Individualistic
LTM	Long Term Memory
MCQ	Multiple Choice Question
OS	Orthopaedic Science
OSCE	Objective Structured Clinical Examination
PAL	Peer Assisted Learning
PBL	Problem Based Learning
PEC	Physical Examination Checklist
PEQ	Post Encounter Clinical Reasoning Questionnaire
PMP	Patient Management Problem
RPC	Reciprocal Peer Coaching
SP	Simulated Patient
STM	Short Term Memory

Chapter 2: Literature Review – Peer Assisted Learning

2.1 Peer Assisted Learning

2.1.1 Definition

Defining peer assisted learning (PAL) is difficult as the literature is rife with definitions describing what is in effect, learners learning from their peers (Lincoln & McAllister, 1993; Martin & Edwards, 1998). Terms such as cooperative/collaborative learning, learning partnerships, peer-centered learning, peer tutoring, reciprocal peer tutoring, peer modeling and peer coaching have all been used to describe PAL methods. In response to this ambiguity in the literature PAL has been defined as a method whereby individuals with equal status actively help and support each other in learning tasks (Topping & Ehly (1998). It is a broad definition and excludes the term cooperative learning. This latter term is seen as a small group learning experience with specific guidelines that guide the learning experience. Topping and Ehly (1998) describe PAL along four dimensions.

Peer assisted learning:

1. represents a group of learning strategies complementary to professional teaching but definitely not surrogate to professional teaching;
2. is structured to ensure gains for all participants in one or more domains;
3. is available to all learners on an equal opportunity basis, since all have something to contribute; and
4. is carefully organised and monitored by professional teachers with an extended conception of their role.

Peer assisted learning, therefore, is used throughout this thesis as a global reference point where peers are learning from their peers. This focus is important in order to manage the breadth of literature on PAL. As a result, this review will focus on three sub-categories of PAL: peer tutoring; peer collaboration; and cooperative learning. Further, this focus will be limited to higher and health sciences education.

The breakdown of PAL into these three categories follows trends in the literature (Damon & Phelps, 1989; Gillies & Ashman, 1995). All three approaches are different in the way they structure learning and in the interpersonal relationships that occur. There are two indices, however, which are common to all three categories: equality and mutuality (Damon & Phelps, 1989). Equality describes the extent to which learners take direction from one another. Mutuality describes the extent to which the learners' discourse is extensive, intimate and connected.

Peer tutoring involves a more able person tutoring a less able person (Damon & Phelps, 1989). The focus of the tutoring is on curriculum content and there are procedures which moderate the interaction (Topping & Ehly, 1998). Peer tutoring, therefore, emulates the traditional teacher-student relationship, although the tutor does not have the same authority or expertise as a professional teacher. This narrower difference in authority and expertise positively influences the instructional discourse because the tutee feels more able to express opinions and ask questions (Damon & Phelps, 1989). Hence, peer tutoring is relatively low on equality but varies on the mutuality scale. Mutuality depends to a large degree on the tutor's interpersonal skills and the tutee's responsiveness to learn (Damon & Phelps, 1989).

Peer collaboration involves two learners working together to solve a task that neither could do previously (Damon & Phelps, 1989; Gillies & Ashman, 1995). Each member begins with similar competencies and learners use each other as sources for mutual discovery, reciprocal feedback and exchanges of ideas. Peer collaboration therefore, is high on the equality and mutuality scale. It simulates discovery learning but puts the discovery in a context of supportive communication and assistance so it is less intimidating for learners (Damon & Phelps, 1989; Sogunro, 1998).

Cooperative learning is an umbrella term which covers a diversity of team learning approaches (Damon & Phelps, 1989; Gillies & Ashman, 1995; Topping & Ehly, 1998). In cooperative learning, teachers organise groups of learners to work through a specific task (Greenwood, Carta, & Kamps, 1990). Heterogeneous groupings of 3 or more students, who differ on various dimensions, are generally the rule. (Watson & Marshall, 1995). Cooperative learning is also predominately used in classroom environments and does not have the same intensity as dyads working together. In

general, it fosters high equality and low to moderate mutuality (Damon & Phelps, 1989). Mutuality is affected by the task sub-division that occurs in cooperative learning. The greater the intra-group discussion, inter-group competition and extrinsic reward, the greater the amount of mutuality.

A more detailed overview of these three PAL methods follows starting with cooperative learning, followed by peer tutoring and peer collaboration.

2.1.2 Cooperative Learning

Cooperative learning was formally introduced by John Dewey in 1899 as a challenge to traditional learning methods which encouraged student competition (Martin & Edwards, 1998). Since then, cooperative learning has become widely implemented in educational settings. Cooperative learning is a good strategy for increasing educational outcomes (Greenwood et al., 1990; Johnson & Johnson, 1978; Johnson, Johnson, & Smith, 1998; Johnson et al., 1981; Larson et al., 1985; Ravenscroft, 1997; Slavin, 1983a; Slavin, 1983b; Slavin, 1987; Slavin, 1990b; Topping & Ehly, 1998). It is also one of the most well researched and accepted curricular innovations currently in practice (Antil, Jenkins, & Wayne, 1998; Johnson et al., 1998).

Specific goal structures need to be in place for cooperative learning to be successful (Johnson & Johnson, 1978; Johnson & Johnson, 1991; Johnson et al., 1981; Slavin, 1983b; Slavin, 1987; Slavin, 1990). These goal structures were originally developed by Deutsch (1949) and influence the cognitive and affective outcomes of the learning experience. These three goal structures are cooperative, competitive, or individualistic in nature. A cooperative goal structure exists when students perceive they can obtain their goals only if the other students with whom they are linked obtain their goals too. This specific goal structure is needed for a successful experience. A competitive goal structure exists when students perceive they can only obtain their goal if the other students fail to achieve their goals. An individualistic goal structure exists if student learning is unrelated to the goal achievement of other students. These latter two goal structures do not encourage cooperative learning.

Evidence for the support of cooperative versus competitive goal structures is seen in a meta-analysis of 46 studies by (Qin, Johnson, & Johnson, 1995). They examined

intra-group cooperation versus individual competition. Cooperative teams generally outperformed individuals competing with one another, with effect sizes ranging from 0.55 to 0.60 reported in the literature. These results were maintained irrespective of age or quality of the study.

On a more specific level, Johnson and Johnson (1991) and Johnson et al. (1998) describe five elements for successful cooperative learning. These are: encouraging learners to interact and work together; positive interdependence; preparation of learners' interpersonal and small group skills; reflection on and evaluation of group processing; and individual accountability. Slavin (1995) in contrast, argues that only two elements are critical: positive interdependence; and individual accountability. Positive interdependence means that group members must work together to gain recognition for their efforts. This can be engendered by giving each student in the group a unique task which is part of the overall group product. Individual accountability means that each member of the group must contribute to the overall process, otherwise, the group cannot be successful in meeting its objectives.

In a review of the literature, Slavin, (1983b) noted that 87.5 per cent of studies which employed group rewards and individual accountability reported statistically significant positive effects on student achievement. Having students merely work together, without these structural factors, did not necessarily lead to enhanced student achievement. Cooperative goal structures also generally lead to superior levels of achievement over competitive and individualistic goal structures (Johnson & Johnson, 1978; Johnson et al., 1981; Slavin, 1990a; Slavin, 1990b). Hence, the structure of the group learning experience is central to the learning outcome.

Cooperative learning principles can be applied to all subject areas and all ages and is particularly useful for tasks that involve concept attainment, verbal problem solving, spatial problem solving, and motor performance (Johnson et al., 1981). The evidence for these claims, however, are stronger at the precollege level (Johnson et al., 1981). The use of cooperative learning without inter-group competition appears to promote higher achievement and productivity than cooperation with inter-group competition, although there are only a few studies supporting this perspective (Johnson et al., 1981). Hence, this argument is still relatively controversial (Slavin, 1977; Slavin,

1995; Topping, 1992). For example, Johnson and Johnson (1987) conducted a meta-analysis of 133 cooperative learning studies that used adults as subjects. Inter-group competition tended to decrease overall productivity. In contrast, a meta-analysis of 72 studies involving cooperative learning in mathematics education found that 50 per cent of the studies with statistically significant effects in favour of the group method employed inter-group competition (Topping, 1992). Differences in these outcomes may be related to the age of the subjects as adults may respond differently to competition than their younger counterparts.

One educational approach that has been used in higher education to promote learning is problem based learning (Barrows & Tamblyn, 1980; Norman & Schmidt, 1992; Schmidt, 1983). This is a cooperative learning technique that has been used rather extensively in health sciences education. It is a subset of peer assisted learning as it uses specific strategies to support learning among peer groups. Problem based learning was initially developed at McMaster University in Ontario, Canada in the late 1960s (Albanese & Mitchell, 1993). Johnson et al. (1998) state that cooperative learning is the core feature of problem based learning. Antil et al. (1998), however, distinguish problem based learning from cooperative learning as the former uses less structure in guiding the activities of the learning group.

Problem based learning is an instructional method using patient problems as a context for students to learn problem solving skills and to acquire knowledge about basic and clinical sciences (Barrows & Tamblyn, 1980; Norman & Schmidt, 1992; Schmidt, 1983). Students are given ambiguous cases to solve in a small group format (five to six students) under the facilitation of a faculty tutor. Advocates of problem based learning state that it is an educational approach that: encourages students to incorporate prior knowledge with new knowledge; places learning within a specific context so it can be more readily transferred to similar situations; and leads to deeper learning because of the elaboration that occurs during the small group discussion (Albanese & Mitchell, 1993; Norman & Schmidt, 1992).

An in-depth review of problem-based learning is beyond the scope of this thesis. It also does not relate directly to the educational model being investigated in this thesis. However, it does employ PAL learning and, therefore, is worthy of review. Several

substantive reviews of the literature have thoroughly investigated this educational approach, largely from the medical education perspective (Albanese & Mitchell, 1993; Berkson, 1993; Kaufman, Portney, & Jette, 1997; Maudsley, 1999; Norman & Schmidt, 1992; Saarinen-Rahiika & Binkley, 1998; Vernon & Blake, 1993).

For the most part, no one has been able to demonstrate conclusively that one curriculum format develops problem solving skills and transfer of medical knowledge better than the other (Berkson, 1993; Kaufman et al., 1997; Norman & Schmidt, 1992; Saarinen-Rahiika & Binkley, 1998; Vernon & Blake, 1993). There are, however, some noteworthy trends in the literature. For example, there is some truth that students in problem based learning programs do not tend to score as well on basic science tests as students from traditional curricula (Albanese & Mitchell, 1993; Vernon & Blake, 1993). Further, students from problem based learning programs have been shown to display more clinical reasoning errors and rely more so on backward reasoning strategies when compared to students from traditional curricula (Saarinen-Rahiika & Binkley, 1998). However, longer term retrieval of knowledge may be more prominent in graduates from problem based programs (Norman & Schmidt, 1992). In contrast, students in problem based learning programs tend to score higher on clinical science tests and on clinical evaluation ratings (Albanese & Mitchell, 1993; Vernon & Blake, 1993). They also appear to be better able to transfer their learning from the academic environment to the clinical setting (Norman & Schmidt, 1992).

Another difference is that students in problem based learning programs appear to study for understanding rather than memorisation and are more self-directed in their learning (Albanese & Mitchell, 1993; Norman & Schmidt, 1992; Vernon & Blake, 1993). Berkson (1993), however, notes that this self-directedness may have more to do with available resources, peer expectations, role modeling and time constraints. Students in problem based learning programs also tend to rate their programs in a more positive light than students from traditional programs and have less learner-centered anxiety (Albanese & Mitchell, 1993; Vernon & Blake, 1993).

Kaufman et al. (1997) compared the clinical performance of physical therapy students enrolled in a problem based and traditional learning program. Overall, there

were no statistically significant differences between the two groups. The researchers note that differences in the two methods of instruction may not have been sufficient enough to produce changes. Approximately 40 per cent of the problem based curriculum actually utilised specific problem based teaching strategies. The remaining 60 per cent of the program was similar in nature to the traditional program. Further, the ability of the clinical evaluation form to actually discriminate high versus low levels of performance was questionable.

Norman & Schmidt (1992) describe a Dutch study of physiotherapy students who were randomly assigned to a muscle physiology course that was taught either by problem based or traditional methods. While students in the problem based course had statistically significant poorer results on a multiple choice test, which was administered immediately after the unit, they recalled up to five times more information six months later on a free recall test of knowledge. One reason for the poor multiple choice test is that students appeared to have studied more highly relevant material which was not assessed on the examination.

Virtually all of the reviews that have compared traditional and problem based learning note several limitations in the research. First, a true distinction between problem based learning and traditional curricula is frequently blurred because of an overlap of these methods in many parts of the curriculum (Berkson, 1993; Kaufman et al., 1997; Maudsley, 1999). Negative outcomes may also not get published because of the political implications for a program. Hence, they may not become part of the body of literature evaluating curricula (Vernon & Blake, 1993). Admission procedures also differ among programs and there may be self-selection of students into the different programs which influences the generalisability of research findings (Berkson, 1993; Saarinen-Rahiika & Binkley, 1998; Vernon & Blake, 1993). Many of the evaluation instruments used to differentiate programs are also insensitive and incapable of capturing good measures of competence (Berkson, 1993; Saarinen-Rahiika & Binkley, 1998; Vernon & Blake, 1993). Hence, it is difficult to ascertain whether research conclusions are idiosyncratic to specific educational programs or are more generalisable features of problem based learning.

2.1.3 Peer Tutoring

Peer tutoring is also a PAL strategy that focuses on learning between pairs. The first systematic use of peer tutoring is associated with Andrew Bell, who in 1789 grouped students by achievement level and then paired them into tutor and tutee roles (McLaughlin, Mecham, & Montague, 1995). Since then, peer tutoring strategies have been implemented on a large scale in elementary and high schools and more recently in higher education where concerns about maintaining quality teaching with less resources is becoming more of an issue (Iwasiw & Goldenberg, 1993; Topping, 1996).

The pairs that are structured in peer tutoring arrangements can be cross age/cross ability, same age/cross ability or same age/same ability (Goldschmid & Goldschmid, 1976; Greenwood et al., 1990; Topping, 1992; Topping, 1996). King (1997), however, sees cross ability or cross age tutoring models as straightforward tutoring since the term 'peer' implies equal status and ability. Peer tutoring, therefore, should be confined to same age/same ability models where one person plays the role of the tutor and the other the tutee. On the other hand, the core feature of peer tutoring arrangements, regardless of age or ability differentials, is students teaching students. Students, broadly speaking, are peers and hence the use of the term, 'peer tutoring'.

Goldschmid and Goldschmid (1976) describe four benefits of peer tutoring: sociopsychological; pedagogical; economic and political. The sociopsychological benefits relate to the support that peers can offer one another during a learning experience. Pedagogical benefits are increased motivation, self-esteem and confidence, and enhanced understanding and cognition. Peer tutoring is also not a costly initiative so it has economic benefits for educational environments with high student:staff ratios. Peer tutoring arrangements also help students adjust to the political realities of being a student in a large organisation.

The contemporary literature is very supportive of peer tutoring as a means of increasing achievement, higher order cognitive skills, and social interaction - with benefits accruing to both members of the pair (Fantuzzo, Riggio, Connelly, & Dimeff, 1989; Goldschmid & Goldschmid, 1976; Greenwood et al., 1990; Lake, 1999; Sharpley & Sharpley, 1981; Topping, 1992; Topping, 1996; Vaidya, 1994).

The literature is fairly consistent in illustrating that gains in achievement using peer tutoring are at least equal to, or better than, instruction by faculty.

While peer tutoring positively benefits both parties, it is usually the tutor who receives greater cognitive gains from the tutoring process (Fantuzzo, 1989; Fantuzzo, Riggio, Connelly, & Dimeff, 1989; Gillies & Ashman, 1995; Griffin & Griffin, 1997; Griffin & Griffin, 1998; Sharpley & Sharpley, 1981; Topping, 1992; Topping, 1996). This occurs because the tutor must re-organise and explain the material in simple terms to the tutee. In doing so, this leads to a better understanding of the material by the tutor (Griffin & Griffin, 1997; Sharpley & Sharpley, 1981; Topping, 1992; Topping, 1996).

Several factors have been shown to influence peer tutoring outcomes. These are: length of the tutoring program; amount of structure; method of pairing tutors and tutees; and tutor training (Sharpley & Sharpley, 1981; Topping, 1992; Topping, 1996). Both low and high achieving students appear to benefit from being tutors and tutees and the greater the structure (i.e., clear guidelines, rules, procedures) the better appears to be the outcome.

Peer tutoring support systems have also been shown to mitigate stress (Fantuzzo et al., 1989b; Topping, 1996). Research using college students has shown that lack of social support and social isolation are related to academic stress and dissatisfaction (Fantuzzo et al., 1989). The peer learning structure appears to empower students (Topping, 1996). Quite often, students do not maximise their skill development in traditional supervisory systems because of the power difference inherent in these arrangements (Topping, 1996). Hence, peer tutoring systems can be very effective in managing learning stress and anxiety.

There are, of course, disadvantages associated with peer tutoring. One disadvantage is that tutors do not have the same mastery of content that a professional teacher possesses (Topping, 1996). Quality control can also be difficult (Maheady, 1998). For example, peers may model behaviours or skills incorrectly (Schunk, 1998; Triggs-Nemshick & Shepard, 1996). Further, they may give inaccurate explanations (Webb, 1982). Learners may also withdraw to work independently, take control or

become the lead problem solver (Damon & Phelps, 1989; Garrett, 1998). Dominance, resentment, jealousy and defensiveness can also develop, interfering with the equality and mutuality of the group (Damon & Phelps, 1989; Watson & Marshall, 1995).

In almost all cases, the importance of preparing or training tutors for their role is emphasised in the literature (Beckon, 1991; Carroll, 1996; Chapman, 1998; Damon, 1984; Fantuzzo et al., 1989; Gerace & Sibilano, 1984; Goldenberg & Iwasiw, 1992; Goldschmid & Goldschmid, 1976; Johnson & Johnson, 1987; King, 1997; Maheady, 1998; Sharpley & Sharpley, 1981; Topping, 1992). Students need to understand basic group processes such as leadership, conflict management and decision making (Cinelli, Wolford-Symons, Bechtel, & Rose-Colley, 1994). An understanding of adult learning practices and the link between peer learning and professional practice should also be highlighted (Goldenberg & Iwasiw, 1992; Lincoln & McAllister, 1993).

2.1.4 Peer Collaboration

Peer collaboration was described earlier as a learning strategy that involves two learners working together to solve a task that neither could do previously (Damon & Phelps, 1989; Gillies & Ashman, 1995). The key point is that each member begins with similar competencies and use each other as sources for mutual discovery, reciprocal feedback and exchanges of ideas. The interaction is more collaborative and is different from traditional tutoring as there is reciprocity of the tutor/tutee roles. Johnson et al. (1998) have referred to collaboration in an educational learning context as 'natural learning' in which students work together in relatively unstructured groups and create their own learning situation.

Reciprocal peer tutoring/learning is one potential form of peer collaboration and involves both parties playing the tutee and tutor role (Cohen & Sampson, 1999; Fantuzzo, 1989; Griffin & Griffin, 1998; Riggio, Fantuzzo, Connelly, & Dimeff, 1991; Riggio, Whatley, & Neale, 1994). Ability levels tend to be similar. Reciprocal peer tutoring, therefore, is quite distinct from peer tutoring and cooperative learning in its structure and organisation. As this learning model is central to the research

questions in this thesis, this review will focus specifically on this particular form of peer collaboration.

A variety of comprehensive reviews and reports on reciprocal peer tutoring are available in the literature (Cohen & Sampson, 1999; Griffin & Griffin, 1997; Griffin & Griffin, 1998; King, 1997; Riggio et al., 1991; Riggio et al., 1994; Topping, 1992; Topping, 1996). These individuals report that reciprocal peer tutoring has been shown to be effective at the undergraduate level in increasing academic achievement and reducing psychological stress. It has also been linked to positive course satisfaction. While there is still a dearth of literature comparing reciprocal peer tutoring to other forms of traditional or PAL models, the literature that is available suggests that highly structured, reciprocal peer tutoring leads to greater achievement, higher satisfaction and less stress than control groups (Hill, Gay, & Topping, 1998).

King (1997) describes a reciprocal peer tutoring program entitled “ASK to THINK - TEL WHY”. While it is described in the context of year seven students, it is reportedly applicable to any learner beyond grade four. This program uses structured categories of open ended questions to guide the tutor and tutee which are designed to scaffold learning and engender metacognition. Students are encouraged to alternate in their role as tutor and tutee to minimise status differentials and to engender a sense of mutual interdependence. This role reversal also gives both parties an opportunity to be in control and promotes a greater sense of self-efficacy and empowerment in the learning situation.

There are, like any other learning strategy, disadvantages to reciprocal peer tutoring (King, 1997). Given that learners are peers, there may be social influences which prevent them from correcting or critically evaluating each others’ arguments. On a cognitive level, given that both learners have roughly similar levels of knowledge/competence, they may make inferences about what the other party knows and omit information incorrectly. In terms of using this model in patient care activities, there may be some confusion about roles given that students are encouraged to alternate between tutor and tutee (Wagstaff, 1989). Many of the disadvantages described earlier for peer tutoring, may equally apply to peer collaboration models.

The use of reciprocal peer tutoring may be extremely beneficial for health science students. However, because patient situations are unpredictable and require a certain element of continuity it may be difficult for the learners to prepare for these situations completely. Hence, the ability of peers to alternate in their role as tutor and tutee equally may be limited. Further, certain aspects of the patient encounter may need to be carried out by one or both of the learners. Reciprocity may also differ depending upon who will be responsible for the ongoing management of the client. As a result, rather than using the term reciprocal peer tutoring, which denotes a balance of operating in the tutor/tutee role, reciprocal peer coaching may be a more appropriate term. Coaching can still be reciprocal but may also be skewed depending upon the clinical and patient care circumstances. For this reason, the term reciprocal peer coaching (RPC) will be used in this thesis to describe the peer collaboration that takes place between peers in a health science clinical education setting.

2.2 Peer Coaching and Professional Development

A considerable body of literature exists in the teacher education field with peer coaching being cited as an important and successful part of professional development programs (Ackland, 1991; Flynn, Bedinghaus, Snyder, & Hekelman, 1994; Hekelman et al., 1994; Johnson & Johnson, 1987; Joyce & Showers, 1995; Kohler, McCullough Crilley, & Shearer, 1997; Martin & Double, 1998; Showers, 1984; Showers, 1985; Williamson & Russell, 1990; Wynn & Kromrey, 1999). While this body of literature tends to focus on teachers and post-graduate professional development, there are many parallels to the use of RPC in undergraduate health science education. This section examines this parallel in more detail and further explores the idea of coaching as a PAL technique.

A variety of terms in the education literature are used to describe coaching: technical coaching, collegial coaching, challenge coaching; team coaching; cognitive coaching and peer coaching (Ackland, 1991; Joyce & Showers, 1995; Wynn & Kromrey, 1999). Ackland (1991) has defined peer coaching as a process where teams of teachers regularly observe one another and provide support, companionship, feedback and assistance. Wynn & Kromrey (1999) define peer coaching as pairs of practicum students, student teachers, or classroom teachers observing each other and

providing consultative assistance in correctly applying teaching skills and proposing alternative solutions to recognised instructional needs. Technical, team and peer coaching have in common a concern for learning and implementing innovations in curriculum whereas collegial and cognitive coaching appear to aim more at improving existing practice. The objectives of all of these coaching sub-categories are highly relevant to health science clinical education programs.

All of these coaching models, except for peer coaching, use 'feedback' to improve current practice. Joyce and Showers (1995) state that feedback is difficult to administer in a peer coaching team as it often becomes evaluative and influences the coaching experience. It also requires a lot of training. Hence they recently omitted feedback as part of their peer coaching model. The omission of feedback during peer coaching is an interesting yet confusing point. How does a coach guide his or her team without providing feedback? Discussion and reflection are key components of a learning experience so how can they not be part of the peer coaching process? The key it would seem is to keep comments, or feedback for lack of a better term, non-evaluative (Ackland, 1991; Showers, 1984; Skinner & Welch, 1996). In other words, providing assistive comments that are not judgmental. By keeping the discussion non-evaluative, that is, confining it to information about the execution of the relevant skills, the integrity of the peer coaching experience is maintained. This integrity is important as the coaching process is highly influenced by the social and psychological aspects of the relationship (Ackland, 1991).

In several reviews of the literature, it has been reported that when theory, demonstration, practice and non-evaluative feedback are combined with coaching, statistically significant gains in performance are achieved (Ackland, 1991; Joyce & Showers, 1995; Skinner & Welch, 1996). Through peer coaching these researchers see learners achieving: knowledge or awareness of current theories and practices; changes in attitudes towards oneself, others and academic content; development of new skills; and the transfer of these new skills so that they are used consistently in practice. Showers (1984, 1985) also notes that coaching facilitates transfer of training. This occurs by learners practicing their newly acquired strategies more frequently and appropriately with their coach. The knowledge associated with these

new strategies is also retained longer because of the rehearsal that takes place. Learners, therefore, are more likely to use these new strategies in the future when faced with a similar context or situation.

This transfer of training is particularly important in the health sciences where biomedical knowledge and psychomotor skills learned in the classroom need to be transferred to the clinical setting. Procedural learning, or putting knowledge and skills into practice, relies heavily on receiving feedback about performance (Johnson & Johnson, 1987). By using non-evaluative feedback in coaching situations, the learner can modify his/her actions until errors are eliminated.

Showers (1984, 1985) describes the benefits of coaching from a teacher development perspective. While educationally focussed, these benefits appear to be easily transferable to other professional disciplines. Coaching is appropriate in situations that require new ways of thinking about objectives. Coaching teams study the rationale behind new skills, see them demonstrated, practice them and learn to provide technical feedback to one another as they experiment with the skill. The feedback that occurs should be accurate, specific and non-evaluative. Coaching also works best when there is not a power differential. Where there is an imbalance in power, it is more difficult to experiment and fail because of the repercussions that may ensue. By placing coaching into the hands of peers, Showers (1985) argues that status and power differentials are minimised and learning is enhanced. These outcomes are difficult to achieve in traditional supervisory systems.

2.2.1 Matching Students for Reciprocal Peer Collaboration

Given that student pairs in reciprocal peer tutoring are supposedly similar in terms of their knowledge or competency, it is important to note whether certain combinations promote or hinder educational outcomes. Greenwood et al. (1990), for example, argue that heterogeneous pairing appears better as one of the pair is more likely to have the answer or abilities to work through the problem. However, Goldschmid and Goldschmid (1976) argue that there are so many complex factors at play in a learning dyad that it is difficult to ascertain with any certainty the best pairing arrangement. Hall et al. (1988), for example, studied 303 freshman and sophomore students in a

learning and study strategies program. They found that extroverts and people who were high in social orientation benefited more from the dyadic learning environment. Cooperative learning situations may arouse the extrovert to an appropriate level for learning but may arouse introverts to a detrimentally high level which interferes with their learning (Berkson, 1993; Goldschmid & Goldschmid, 1976; Hall et al., 1988; Topping, 1996). Further, extroverts tend to do better than introverts in peer tutoring, suggesting that an interaction between learning method and personality/learning style may also have an influence.

In investigating the issue of matching students further, Riggio et al. (1994) evaluated 206 undergraduate students enrolled in two psychology units. A 25 item MCQ covering material from these two courses was administered at the beginning and end of the units. In this study, students were organised into four groups with the intent of evaluating which type of dyadic pairings led to the highest cognitive gains. The students' self-reported grade point average was used to form dyads of low:low, medium:medium, high:low and high:high students. No statistically significant differences in cognitive gains were noted across the four groups. A comparison was also made of low ability students across the low:high and low:low groupings and high ability students across the high:high and low:high groupings. All students demonstrated cognitive gains. Interestingly, the high:high students had lower cognitive gains than those high students who were paired with low students.

Chapman (1998) notes that there is some evidence that low achievers make greater gains in mixed-ability pairs. However, Chapman warns that the discrepancy in competence should not be so great that it precludes the development of mutual understanding between the pair. Riggio et al. (1994) similarly looked at academic ability as a factor in predicting success in a reciprocal peer tutoring model. They organised low, medium and high achieving adults into various paired combinations. They found that cognitive gains achieved through reciprocal peer tutoring were relatively unaffected by academic level. They explain this finding by the fact that both parties must play the role of tutor and tutee. Hence, they must both prepare and teach their partner which in itself leads to cognitive gains.

2.3 Peer Assisted Learning: Theoretical Expositions of its Efficacy

While the structure of cooperative learning, peer tutoring and peer collaboration are different from one another, one common feature they all share is that peers are interacting with, and learning from, their peers. Johnson et al., (1998) describe three theoretical categories that can be applied to cooperative or peer assisted learning at the college level. They are social interdependence theory, cognitive development theory and behavioural learning theory.

2.3.1 Social Interdependence Theory

This theory builds upon the work of Deutsch (1949). The basic premise of this theory is that the way social interdependence is structured determines how individuals interact, which in turn influences outcome (Johnson et al., 1998). Johnson et al. (1998) describe three goal structures which form the basis of social interdependence theory. These were presented earlier and are highlighted here to illustrate this theory. Positive goal structures lead to promotive interaction which encourage learners to facilitate each others' learning. Competitive goal structures lead to oppositional interaction as learners discourage and obstruct each other's efforts to learn. In individualistic goal structures, there is no interaction between learners.

2.3.2 Cognitive-Developmental Theory

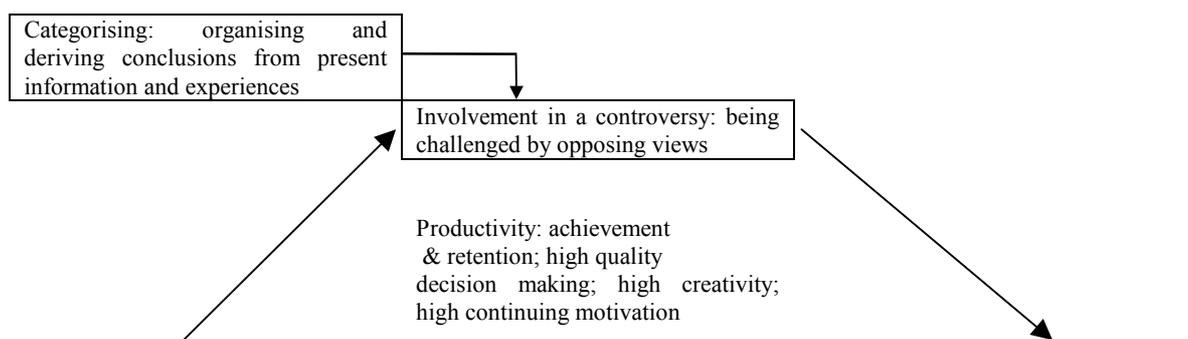
This theory sees cooperation among peers as an essential pre-requisite for cognitive growth and flows from Piaget and Vygotsky and their perspectives on socio-cognitive conflict (Johnson et al., 1998). Most of the literature supporting cooperative learning and PAL appears to fall within this theoretical category.

Damon (1984) and Foot and Howe (1998) provide an excellent theoretical review of PAL from the cognitive-developmental perspective. They incorporate the theories of several well known psychologists (Piaget, 1977; Sullivan, 1953; Vygotsky, 1978; Vygotsky, 1986). In their review they state that peer interaction is seen to promote cognitive development by creating critical cognitive conflict. If the learner is aware of a contradiction in their knowledge base, the experience creates a disequilibrating effect, which instigates the learner to question his or her beliefs and to try out new ones. Peers are a compelling source of conflict because they speak on levels which

can be easily understood by one another (Damon, 1984; Foot & Howe, 1998; King, 1997). Further, the informational communications between peers are less threatening than the corrective advice from a superior. Where there are large ability differentials, however, status differences may be so great that they interfere with learning. Less able counterparts may lessen their interaction because of the potential embarrassment they may experience from making mistakes.

Several researchers have expanded upon Piaget's developmental perspective (Johnson, 1981; Johnson & Johnson, 1978; Johnson, Johnson, & Smith, 1986; King, 1997; Slavin, 1987). They argue that one of the benefits of cooperative learning is controversy. The intellectual disagreements that occur in socio-cognitive learning situations create conceptual conflicts that motivate learners to seek out new information. When managed properly, this 'structured controversy' can lead to higher achievement levels (Johnson, 1981; Johnson et al., 1986).

There are two aspects to this controversy (Johnson, 1981). The first involves conceptual conflict which occurs when two ideas do not appear compatible. The second aspect involves epistemic curiosity which is an active search for more information. Both of these features increase in their utility when the disagreement is larger and more frequent. The controversy promotes higher cognitive and moral reasoning because there are increases in perspective taking and reductions in egocentric reasoning (Johnson, 1981). To promote controversy, Johnson (1981) recommends heterogeneity in the group such as age, gender and ability level differences. Johnson et al. (1986) illustrate the process of controversy diagrammatically (Figure 2.1).



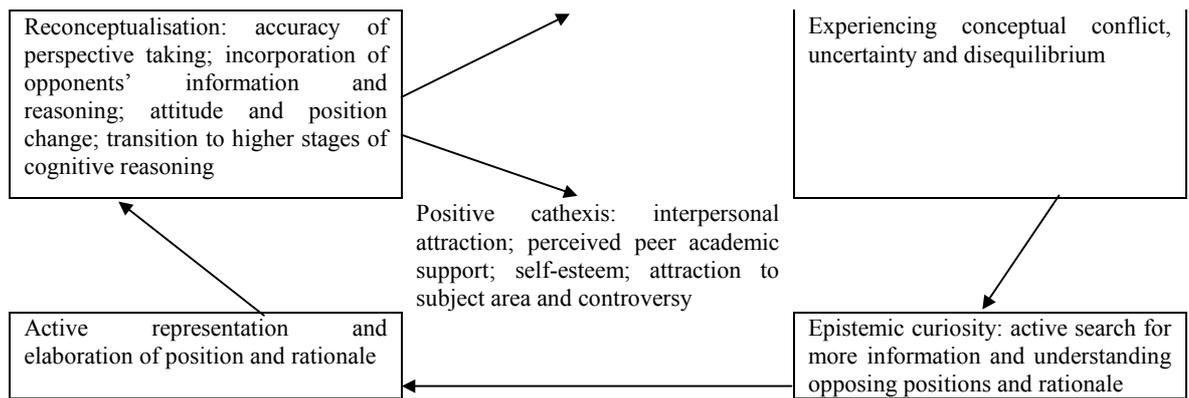


Figure 2.1: The Process of Controversy
 (Source: Johnson et. al., 1986)

These same theoretical principles are described by Joyce and Weil (1996). They report that the synergy that is created in a peer centered learning experience generates more motivation to learn than in competitive or individual learning efforts. The social learning activity that is generated creates more intellectual activity than solitary study since it requires peers to reconstruct their knowledge based upon the deliberations they experience with their peers (Joyce & Weil, 1996; King, 1997). The conflicts that emerge energise learners to pursue the problem further, which leads to new and emerging patterns of knowledge and skill.

Transfer of training is also facilitated by this socio-cognitive learning approach (Bandura, 1971; Bandura, 1997). Joyce and Showers (1982) have noted that an adequate background in theory, demonstration and practice with feedback does not necessarily translate into appropriate performance in context. Working alongside a peer is one mechanism of promoting this transfer as it provides companionship and support, technical feedback and analysis of executive control functions. Schunk (1998) calls this ‘peer modeling’ and has demonstrated how this technique can be used to promote transfer of skills, attitudes, beliefs and behaviours.

Damon (1984) and Foot and Howe (1998) also provide an overview of the theories of Vygotsky (1978,1986) and his social interactionist view of cognitive development. Vygotsky reports that peers benefit from one another by internalising the cognitive processes implicit in their interactions and communications. The peer dialogue that results emulates several critical features of rational thinking. In particular: the verification of ideas; the planning of strategies; the symbolic representation of

intellectual acts; and the generation of new solutions. Further, Vygotsky argues that the social and cognitive interaction with a more capable peer allows the less capable learner to enter new areas of potential. Vygotsky calls this the 'zone of proximal development'.

The peer dialogue that occurs in these learning situations is also central to learning outcomes. O'Donnell and Topping (1998) state that peers provide greater volumes of feedback which is generally much more immediate than the feedback that comes from supervisors. The volume and immediacy of feedback also makes up for any differences in the quality of feedback that would normally come from more informed supervisors. This feedback influences the development of new schemata through processes of cognitive restructuring, metacognitive awareness and post hoc reflection (O'Donnell & Topping, 1998).

2.3.3 Behavioural Learning Theory

This final theoretical perspective states that learners will work hard on those tasks for which they secure a reward and will fail to work on tasks that yield no reward or yield punishment (Johnson et al., 1998). This behavioural perspective originates from the work of B.F. Skinner and his work on operant conditioning (Biehler & Snowman, 1997). His principle thesis was that individuals operate on their environment in order to obtain or avoid particular consequences. Through either positive or negative reinforcement, or punishment, certain behaviours can be promoted or extinguished. Although this represents a fairly simple perspective, it opened the way for future theorists to expand upon their views of behaviour and learning.

The social learning theories of Bandura (1971, 1997) and his views on group learning, provide a useful framework for examining the influence of PAL on learner efficacy. Bandura describes three kinds of reinforcements that influence learning outcomes. The first is direct external reinforcement. Under this form of reinforcement, persons regulate their behaviour on the basis of the consequences they experience directly. The second is vicarious reinforcement. This type of reinforcement occurs by observing the experiences of others and then modifying your

own behaviour based upon the consequences you have just observed. Self-administered reinforcement involves regulating one's own behaviours according to standards. Reciprocal peer coaching, or the peer modeling approach described by Schunk, (1998) provides rich opportunities for these three types of reinforcement to occur. For example, feedback from a peer or observing a peer may help a novice clinician recognize certain consequences of their behaviour. All of these potential reinforcements, of course, contribute to the learners' metacognitive learning framework as they provide opportunities for identifying knowledge gaps and deficiencies. Hence, there is still a cognitive-developmental perspective to Bandura's social learning framework.

Slavin (1983b, 1987, 1991) also adheres to the behavioural learning perspective but also integrates his views on behavioural learning with social interdependence theory. Advocates of the motivational or behavioural learning perspective concern themselves with the reward structures of cooperation, competition and individual effort. In cooperative goal structures, individual members must help others if their group is to succeed or avoid punishment. Slavin (1983b,1987) notes that studies which do not incorporate the use of reward structures in their methods, and just have students working together, do not show statistically significant achievement differences between experimental and control groups. Although this seems to be the norm, not all well controlled studies on cooperative learning, which have incorporated incentive structures, lead to increases in achievement (Watson & Marshall, 1995). Hence, this issue of rewards remains somewhat controversial.

2.4 Peer Assisted Learning In Higher Education

This chapter so far has focussed on defining PAL and providing an overview of the theoretical reasons behind this learning strategy. The remainder of this chapter now looks at the more practical side of PAL in higher education and health sciences education. High quality research on PAL at the tertiary level is quite sparse in comparison to the breadth of literature at the elementary and secondary school level (Hampton & Grudnitski, 1996; Martin & Edwards, 1998; Milson & Laatsch, 1996; Slavin, 1990; Smith, Hinckley, & Volk, 1991; Topping, 1996). At the pre-tertiary level, most of the research illustrates that cooperative learning is more effective than

competitive or individual learning in increasing achievement (Hampton & Grudnitski, 1996; Johnson & Johnson, 1978; Johnson et al., 1981; Slavin, 1983b; Slavin, 1987; Slavin, 1990).

At the tertiary level, however, there is not the same degree of consensus that cooperative learning benefits all learners (Hampton & Grudnitski, 1996; Slavin, 1990). However, most studies and reviews have demonstrated that increased academic achievement in tertiary level students using cooperative learning strategies is highly likely (Cook, 1991; Goldschmid & Goldschmid, 1976; Hampton & Grudnitski, 1996; Holt, Michael, & Godfrey, 1997; Johnson et al., 1998; Milson & Laatsch, 1996; Riggio et al., 1991; Riggio et al., 1994; Topping, 1996). Many of the studies reported in these reviews, however, are descriptive and rarely use experimental methods which incorporate control groups (Hill et al., 1998; Lindquist, 1997; MacFarlane & Joughin, 1994; Rizzolo, 1982; Sullivan, 1996; Topping et al., 1997).

Johnson et al. (1998), however, comment on over 300 studies which have examined cooperative learning, individual learning and competitive learning situations in college and adult education settings. With respect to academic success, their meta-analyses of 168 studies over the past 70 years demonstrate that cooperative learning methods promote higher achievement than competitive approaches and individual approaches with effect size differences of 0.49 and 0.53 respectively. These effect sizes describe significant increases in achievement in the areas of knowledge acquisition, retention, accuracy, creativity in problem solving and higher level reasoning. The quality of social support in the learning relationship is also greater in cooperative learning than in competitive learning and individualistic learning with effect sizes of 0.60 and 0.51 respectively. Cooperative learning is also able to promote higher self-esteem than competitive learning and individual learning with effect sizes of 0.47 and 0.29 respectively.

With respect to specific studies that have investigated reciprocal peer tutoring/coaching arrangements, several are reported in the literature. Goldschmid & Goldschmid (1976) were one of the first to investigate reciprocal peer learning arrangements in higher education. They used a strategy called the “learning cell”.

Each learning cell was composed of two students who each alternated in their roles as teacher and student. This method was applied to 250 undergraduate psychology students in a reading comprehension and retention test. Students in the learning cell had statistically significant higher examination scores and rated the learning experience more highly than students who were engaged in the task independently.

Riggio et al. (1991, 1994) have investigated the influence of reciprocal peer tutoring on student achievement in great detail. Riggio et al (1991) evaluated 85 undergraduate students enrolled in a psychology unit. Cooperative pairs had statistically significant higher scores than the individual group on the unit examination. Dyads who were given a formalized structure to follow showed the highest gains in academic achievement and psychosocial adjustment. This same group also had statistically significant higher course satisfaction scores.

Fantuzzo (1989) and Fantuzzo et al. (1989) have also used reciprocal peer tutoring at the tertiary level. They found that this strategy results in greater cognitive gains, lower levels of subjective distress and higher course satisfaction. Fantuzzo et al. (1989) specifically compared structured and unstructured reciprocal peer tutoring to structured and unstructured individualistic learning using 100 psychology students enrolled in a psychology course. A variety of instruments were used to evaluate knowledge, fear of evaluation, distress and depression, and student satisfaction. Structured groups were required to generate a 10 item test before each class unit exam, complete with answers. In the case of the structured reciprocal peer tutoring group, the tests were administered to their partners and evaluated collaboratively. In terms of the study's outcomes, the dyadic tutoring arrangements and the structured groups achieved the highest scores. A similar study was also conducted by Fantuzzo (1989) using 49 students enrolled in another psychology unit. Again, it was the structured dyadic group that had statistically significant higher knowledge scores than the other dyadic/unstructured and individualistic/unstructured groups. The dyadic structured group also had statistically significant lower distress scores and statistically significant higher satisfaction ratings for the course.

Griffin & Griffin (1997) adopted methods similar to Fantuzzo and colleagues in two studies involving 93 and 38 post-graduate students respectively who were enrolled in

an educational research unit. Unit test results, as well as inventories evaluating test anxiety and self-efficacy were used to compare the results for the reciprocal peer tutoring and individualistic groups. While there were no statistically significant differences in test scores between the groups, there were positive effect size differences in favour of the reciprocal peer tutoring group. The largest effect sizes were seen in test scores that related to higher order cognitive outcomes. There were no differences between the groups on measures of anxiety or self-efficacy. Griffin & Griffin (1997) note that their methods may have not provided students with enough opportunity for face to face interaction, particularly in the first experiment which was a very episodic experience. This may have minimised the magnitude of the results and the ability of the peer groups to effect changes in their performance.

Not every study that has employed PAL has lead to greater educational achievement in the higher education sector (Hertzog & Lieble, 1996; Lemke & Basile, 1997). Cotton and Cook (1982) also refute the results of the meta-analysis conducted by Johnson et al. (1981) and state that neither cooperative, competitive or individual reward systems are superior in promoting educational advancement and productivity. Cotton and Cook (1982) feel there are situational factors which influence cooperative learning which were not addressed adequately by Johnson et al. (1981). For example, several studies that were used in the meta-analysis were flawed in that they did not examine the effects of interactions. Anderson, Reder, and Simon (1996) and Wood and O'Malley (1996) also note that many of the studies which support the benefits of cooperative learning between peers have been experimentally based and/or poorly controlled. Many of these benefits have not actually surfaced in real classroom practice where class sizes are often quite large. Wood and O'Malley (1996) postulate that many of the conditions that are created in the laboratory are difficult to create in the classroom. For example, instructors must: understand what students know and do not know; be confident that the task structure they have selected for the group is appropriate; be able to monitor groups; and be prepared to assist students if they lack the appropriate skills to scaffold learning interactions. This may explain why many studies fail to demonstrate statistically significant differences in educational outcomes between cooperative and individualistic groups.

The small proportion of studies that do not report educational advantages when PAL is employed, however, are in the minority. Johnson et al. (1998) has stated that there is little doubt that cooperative learning is appropriate to higher education.

2.4.1 Peer Assisted Learning in the Health Sciences

There is good support for PAL in health sciences education (Costello, 1989; DeClute & Ladyshefsky, 1993; Gandy, 1999; Goldenberg & Iwasiw, 1992; Iwasiw & Goldenberg, 1993; Kleffner & Dadian, 1997; Ladyshefsky, Barrie, & Drake, 1998; Lynch, 1984; Mattana, Shepherd, & Knight, 1997; Milson & Laatsch, 1996; Trautwein, Racke, & Hillman, 1996). For the most part, these research and descriptive reports suggest that PAL strategies enable students to perform at higher cognitive levels and to transfer learning to new situations. Many of these claims, however, need to be tested with much more rigour (Iwasiw & Goldenberg, 1993; Lemke & Basile, 1997; Lincoln & McAllister, 1993). The lack of well controlled studies comparing PAL to traditional forms of learning in the clinical education environment support this claim. A few studies do exist, however, and clearly provide support for PAL in clinical education (DeClute & Ladyshefsky, 1993; Iwasiw & Goldenberg, 1993; Lynch, 1984; Mattana et al., 1997; Milson & Laatsch, 1996; Trautwein et al., 1996).

DeClute and Ladyshefsky (1993) studied the clinical competency outcomes of 108 final year physiotherapy students in individual and cooperative (dyads) learning placements. Using a reliable and valid standardised evaluation instrument and an experimental and control group, it was found that the higher clinical competency scores of the students in the cooperative group were statistically significant in comparison to the students in individualistic models. Competencies requiring higher levels of judgment and reasoning were particularly enhanced by the cooperative learning experience.

Iwasiw and Goldenberg (1993) describe the results of a well controlled study using nursing students in the clinical setting. Students were required to complete a surgical dressing procedure. Two groups were created. The first was a peer supervision group in which both parties coached each other through the procedure. The other was the

more traditional approach involving a student on their own with episodic supervision by a clinical instructor. Both cognitive and psychomotor gains were evaluated. The peer group obtained statistically significant higher cognitive scores for the procedure. While the psychomotor and time to complete task measures improved more so for the peer group, these changes were not statistically significant.

Milson and Laatsch (1996) studied the performance of students enrolled in a laboratory instrumentation course. The examination scores of nine consecutive classes were evaluated over the course of nine years using a variety of individualistic and cooperative learning techniques. Three types of examination scores were collected: individual student outcome scores (solo); group based outcome scores (group); and a mixture of both (mix). While there is little description of the cooperative learning strategies actually employed, the group and mix scoring methods had statistically significant higher scores on several of the continuous evaluation tests in comparison with the solo group. Final examination scores were highest for the group (78.4 per cent), followed by the mix (75.4 per cent) and solo (73.9 per cent) although these were not statistically significant different across the groups.

By far, the majority of reports describing PAL in the health sciences literature are descriptive in nature (Aviram, Ophir, Raviv, & Shiloah, 1998; Beeken, 1991; Costello, 1989; DeDea, 1996; Gandy, 1999; Haffner-Zavadak, Konecky-Dolnack, Polich, & Van Volkenburg, 1995; Kleffner & Dadian, 1997; Ladyshevsky, 1993; Ladyshevsky, Barrie, & Drake 1998; Martin & Edwards, 1998; Triggs-Nemshick & Shepard, 1996; Wagstaff, 1989; Williams, 1995). The general consensus is that PAL can be used to: facilitate the transfer of biomedical knowledge to the clinical setting; develop positive attitudes towards subject matter; reduce anxiety and stress; encourage lifelong learning; and enhance problem solving and critical thinking.

With respect to clinical education, positive PAL experiences have been reported in physical and occupational therapy literature (Haffner-Zavadak et al., 1995; Martin & Edwards, 1998; Triggs-Nemshick & Shepard, 1996)). Through this model of learning, students were able to reduce their stress through mutual support, learned from each other through joint problem solving initiatives and saw more cases

because of the joint collaboration. In some cases though, more advanced students were found to dominate learning situations. Other negative aspects of the paired learning were that on some occasions students would provide incorrect information to one another. Further, students occasionally would leave the problem solving to the other party.

On a conceptual level, Lincoln & McAllister (1993) argue that emotions may be more freely expressed in PAL situations as there is often more empathy and less risk of judgment. Learners can also exercise more independence and control over their learning which is quite the opposite to traditional models where the student is in a dependent role and choices are limited. Kleffner and Dadian (1997) also argue that PAL initiatives be employed when higher order cognitive skills such as concept formation, application, problem solving and inference making need to be developed. These cognitive skills are a necessity in the clinical setting.

There are however, reported challenges associated with the use of PAL in clinical learning situations (Gandy, 1999; Ladyshevsky & Healey, 1990). Some examples are: PAL requires more planning, effort and organisational time by the clinical instructor; there must be a sufficient volume of patients to provide enough learning experiences for the student group; students may compete with one another; and conflict may arise within the group.

2.4.2 The Development of Generic Skills using Peer Assisted Learning

The development of generic skills and behaviours appear to be developed more readily using PAL (Beeken, 1991; Candy & Worrall-Carter, 1999; Hunt & Higgs, 1999). Davis (1998) defines generic skills as attributes, characteristics or behaviours that are not explicitly part of the profession's core of knowledge and technical skills but are nevertheless required for success in the profession. Some of these skills are self-assessment, interpersonal skills, communication skills, ability to give and receive constructive feedback, problem solving, critical thinking, professionalism, and stress management (Davis, 1998; Gandy & Jensen, 1992; Hunt & Higgs, 1999).

The importance of generic skill development has become more prominent in the higher education sector. Antil et al. (1998) note that most instructional strategies

focus on academic goals with few opportunities for students to develop interpersonal skills. Interpersonal skills are extremely important for success as a health care practitioner. Peer assisted learning, with its dual emphasis on academic achievement and interpersonal skill development, allows both objectives to be met. Even if PAL does not lead to increased academic achievement, this learning approach may still be worthwhile if it improves interpersonal relationships, self-esteem and communication skills (Milson & Laatsch, 1996).

The literature is abundant with evidence that PAL leads to the development of generic skills and behaviours in professional programs (Cinelli et al., 1994; Garrett, 1998; Goldenberg & Iwasiw, 1992; Iwasiw & Goldenberg, 1993; Johnson, 1981; Mattana et al., 1997; Sullivan, 1996). The importance of generic skill development is also important from the perspective of the employer. Wilson, Stull, & Vinsonhaler (1996), for example, cite the results of two national surveys undertaken in the USA which describe what employers look for when hiring employees. At the top of the list are attitude, verbal and written communication skills, problem solving skills and ability to work within a team.

Working alongside peers may also promote learners' sense of perceived self-efficacy or confidence (Chapman, 1998; Schunk, 1998). Working alongside peers who become competent over time may also be better for learners than working with peers who are already competent. One reason for this may be that the observer can break down task complexity and model the behaviour based upon their observations of their peer.

Clearly, there are many benefits associated with the use of peer assisted learning. This chapter has attempted to provide an overview of peer assisted learning. Three categories of peer centered learning: cooperative learning, peer tutoring and peer collaboration have been reviewed along with the theoretical reasons for their efficacy. Reciprocal peer coaching, as an instructional method for teaching and learning in the health sciences clinical education environment, appears to hold much promise.

Chapter 3: Literature Review – Simulated Patients

This Chapter presents a review of the literature on simulated patients (SPs) and methods to evaluate clinical performance and problem solving. The first part of this chapter focuses on SPs and how they can be used in teaching, evaluation and research. The subsequent sections of this chapter focus on methods for measuring or evaluating clinical performance and problem solving.

3.1 Simulated Patients: A Definition

To appreciate the breadth of the literature on SPs, the National Board of Medical Examiners in the United States of America listed 209 references as of 1991 (Colliver & Williams, 1993). Tamblyn et al. (1992) also noted that in 1989, seventy percent of medical schools in Canada and the United States of America were using SPs for teaching and evaluation. Six international conferences on teaching and assessing clinical competence have also been held since 1985 with over 200 of the presentations related to simulated patients (Colliver & Williams, 1993; Ferrell, 1995). The American Association of Medical Colleges organised a consensus conference on SPs in 1993 and both *Academic Medicine* and *Teaching and Learning in Medicine* journals have had special issues on simulated patients (Ferrell, 1995).

Barrows and Abrahamson (1964) were the first people to formally develop a procedure for live patient simulations. The people they trained to portray patients were initially referred to as ‘programmed patients’ and were used to appraise medical student performance in clinical neurology. In 1987, a formal publication outlining SP technology and training was produced (Barrows, 1987). Using the terms ‘simulated’ and ‘standardised’, Barrows defined this type of patient as:

“a person who has been carefully coached to simulate an actual patient so accurately that the simulation cannot be detected by a skilled clinician. In performing the simulation, the simulated/standardised patient presents the ‘gestalt’ of the patient being simulated; not just the history, but the body language, the physical findings, and the emotional and personality characteristics as well.” (p.1).

This definition was further refined in 1993 by defining a ‘standardised patient’ as a person who has been coached to present their *own illness* in such a way that it does not vary from student to student (Barrows, 1993). The term ‘simulated patient’ was defined as a normal person who has been carefully coached to accurately portray a specific patient when given the history and physical examination (Barrows, 1993). It is this term, simulated patient (SP), that is used in this thesis.

Numerous techniques are often confused with SP technology such as role play, pseudopatients, surrogate patients, subjects, practical instructors and patient instructors (Barrows, 1987; Barrows, 1993; Frazer & Miller, 1977; Norman, Barrows, Gliva, & Woodward, 1985a; Ostrow, 1980; Owen & Winkler, 1974; Stillman et al, 1983; Stillman, Ruggill, Rutatla, & Sabers, 1980). The reason why these techniques cannot be described as SPs is that they do not use the methods explicitly described by Barrows (1987) for the training of simulated patients.

3.2 Educational Applications and Advantages of Simulated Patients

Simulated patients in their various forms have been used extensively in medicine as well as other disciplines. The use of SPs in instruction has also increased since 1992 to teach communication, history taking and physical examination (Association of American Medical Colleges, 1998). A large number of published articles describe the educational benefits of SPs in health sciences education. These are identified in Table 3.1.

A main advantage cited by many of the authors in Table 3.1 is that the SP can be used to standardise the patient experience. Simulated patient responses are reproducible. This ensures that all students see, physically feel and experience the same relative things. Real patients may change their history or may differ in their reaction across multiple students. This would make it difficult to discuss certain aspects of a patient’s presentation in a tutorial, for example, if the real patient varied considerably across the students.

Table 3.1: Educational Uses of Simulated Patients

Authors	Educational Application
(Ainsworth et al., 1991; Fraxer & Miller, 1977; Stillman, Ruggill & Sabers, 1978) (Barrows & Abrahamson, 1964) (Barrows, 1971; Barrows, 1987; Edwards, Franke, & McGuiness, 1995; Norman et al., 1985a; Swanson & Stillman, 1990) (Barrows, 1993)	<ul style="list-style-type: none"> • Instruction and evaluation of interviewing and physical examination skills for medical students • Appraisal of student performance - general • General overview of educational advantages
(Bowman et al. 1992)	<ul style="list-style-type: none"> • Instruction and evaluation of clinical reasoning • Appraisal of student performance • Instruction in history taking, physical examination, interviewing
(Coggan, Knight, & Davis, 1980)	<ul style="list-style-type: none"> • Continuing education for primary care physicians in sexual health counseling • Evaluation of history, physical examination and interviewing skills with focussed feedback in family medicine
(Ferrell, 1995)	<ul style="list-style-type: none"> • Overview of evaluation procedures using simulated patients
(Foley, Nespoli, & Conde, 1997)	<ul style="list-style-type: none"> • Registered nurses - development of communication skills
(Gerber, Albanese, Brown, & Matthes, 1985)	<ul style="list-style-type: none"> • Instruction of male genito-rectal examination for medical students
(Gold et al., 1995)	<ul style="list-style-type: none"> • Skill development for medical students in a gerontology program - with focussed feedback
(Hanna, 1991)	<ul style="list-style-type: none"> • Overview of the use of patient simulations in nursing education
(Hannay, 1980)	<ul style="list-style-type: none"> • Instruction in interviewing techniques for trainees in general medical practice
(Hasle, Anderson, & Szerlip, 1994)	<ul style="list-style-type: none"> • Overview of the costs and benefits of using simulated patients in educational programs
(Haydon et al., 1994)	<ul style="list-style-type: none"> • Use of simulated patients in examinations to evaluate deficits in the curriculum
(Heaton, Watson, & Alger, 1994)	<ul style="list-style-type: none"> • Use of simulated patients to introduce elements of health risk appraisal and risk factor reduction
(Jenkins, 1991)	<ul style="list-style-type: none"> • Continuing education for general practitioner vocational trainees
(Liu, Schneider, & Miyazaki, 1997)	<ul style="list-style-type: none"> • Instruction of clinical skills to physical therapy and occupational therapy students
(McAvoy, 1988)	<ul style="list-style-type: none"> • Instruction of interviewing, history taking, physical examination, diagnosis, and management for medical students in general practice - with focussed feedback sessions
(Ogden, Barnhart, & Davis, 1987)	<ul style="list-style-type: none"> • Use of an extended SP case in a problem based curriculum.
(Owen & Underwood, 1980) (Sanson-Fisher & Poole, 1980)	<ul style="list-style-type: none"> • Instruction in the doctor-patient relationship • Communication skills training for second year medical students
(Smit & Van Der Molden, 1996)	<ul style="list-style-type: none"> • Evaluation of interviewing and counseling skills in undergraduate students enrolled in a psychology program
(Stillman et al., 1980)	<ul style="list-style-type: none"> • Instruction of phys. exam. & diagnosis to med. students using pulmonary patient instructors

The literature also notes that the SP case can be designed in such a way that it matches curriculum objectives with the level of student training. This provides a rich opportunity for bringing course content alive, since students can apply their learning in a very specific way. Simulated patients can also be used to illustrate changes in patient status over time. These changes can be simulated so students have an opportunity to experience the history of a case from start to finish in an encapsulated time frame.

Simulated patients are also more accessible than real patients and can be used in non-clinical environments such as the classroom or tutorial setting. They are also safe to use with inexperienced students. This eliminates many of the ethical and safety risks associated with the use of real patients. This is particularly useful when the management of emergency or high-risk situations need to be managed prior to their implementation on real patients.

In a review chapter on simulated patients, Edwards et al. (1995) describe several specific educational advantages of simulated patients. These authors state that SPs can be used for a wide variety of skill development in such areas as interviewing, counseling, data gathering, physical examination, psychosocial assessment and clinical decision making. The use of simulations in the educational environment also gives the educator an opportunity to create a more experiential learning environment (Hanna, 1991). For example, social learning influences such as incidental learning, reflective thinking and learner relevance emerge when simulations are used and can lead to increased student competence.

There are, however, some disadvantages associated with SP technology. Simulated patients should not replace exposure to real patients. Another issue is the time it takes faculty to prepare the SP program. Training of the SP can vary although a SP can be trained to perform reliably in as little as three hours (Barrows, 1987). The other potential disadvantage is cost. Simulated patients can be expensive if used extensively in a program. However, Barrows (1993) argues that it is more economical to use a SP to increase the direct observation of students and to teach basic skills than to use faculty members for this purpose. Hasle et al. (1994) also supports the views put forward by Barrows (1993). He found that students who

practiced their physical examination skills on SPs in combination with their hospital based diagnosis sessions did equally as well in an objective structured clinical examination as students in the previous year who only had hospital based diagnosis sessions. The cost savings arose from the former group having eight less hospital based diagnosis sessions. While this result provides some economic support for SP programs, the study by Hasle did not have a control group of students who had eight less hospital visits and no experience with simulated patients. Barrows, (1993) notes that “the cost-effectiveness of a SP program is difficult to estimate because of differences in the scope of applications, environmental factors and pre-existing infrastructures” (p.453).

3.3 Simulated Patients and Evaluation of Clinical Competency

The use of SPs in the assessment of clinical competency and performance is well documented in the medical education literature and to a much lesser extent in the allied health sciences. The use of SPs in medical student assessment has also increased remarkably since 1992 (American Association of Medical Colleges, 1998). In a review of the literature, Vu and Barrows (1994) state that:

“SP technology is useful in the assessment of performance not only in the health professions but also in any profession where it is important to assess the examinees’ abilities to determine *on their own* the necessary tasks to be performed in a situation, to carry out those tasks correctly, and to interact and relate effectively with the individual(s) encountered in the situation” (p. 27).

3.3.1 Fidelity (Content Validity) of Simulated Patients

To achieve the measurement outcomes that Vu and Barrows (1994) describe with any degree of confidence, the SP must possess high fidelity or external validity. Fidelity is defined as, “exactness of reproductive detail” (Avis, 1980) or “accuracy of description, translation, sound reproduction etc...” (Guralnik, 1977). If decisions are going to be made concerning student competence, SPs must be able to portray the history and physical features of a client with high fidelity. The SP must possess acting skill and be able to simulate certain physical findings so they are believable by

the practitioner. A range of findings can be simulated in the neurological, cardiopulmonary and musculoskeletal systems (Barrows, 1993; Reteguiz & Cornel-Avendano, 1999). With respect to the musculoskeletal system, which is the focus of this particular thesis, findings such as costovertebral tenderness, joint restriction, tenderness, muscle spasm, muscle weakness, and sensory losses, for example, can be readily reproduced in a simulation.

Several methods are suggested for ensuring and evaluating SP fidelity. Training is one component. Ainsworth et al. (1991) and Burri, McCaughan, and Barrows, (1976) have suggested several methods to ensure fidelity as part of the training process. These are having the SP performance observed and assessed by clinicians, developing the SP case using experts in the field and having them work closely with the patient trainer. Videotaping the SP performance for feedback and monitoring purposes is also another useful method. Other methods for evaluating SP fidelity are asking certified practitioners and students to rate the realism of the case, measuring the extent to which SPs can gain undetected access into clinical practices and evaluating whether practice changes when simulated versus real patients are used. A variety of studies which have employed these methods are summarised in the following paragraphs.

Bowman et al. (1992) looked at the effect of educational preparation on physician performance with a SP carrying a sexually transmitted disease. In this study, 99 per cent of 232 participants found the case to be believable even though an overt SP was used. Helfer, Black, and Teitelbaum (1975) used real and simulated mothers to evaluate third year medical students' interviewing skills in a paediatric setting. All students improved on the second interview, regardless of whether they encountered a real or simulated mother. Nowotny and Grove (1982) used real and SPs in an objective structured clinical examination. Of the 109 students evaluated, there were no statistically significant differences in student scores as evaluated by either the real or simulated patients. Students were also unable to determine beyond chance which patients were real or simulated.

Other studies have attempted to measure SP fidelity by evaluating detection rates when covert SPs gain access to a medical practice. McClure et al (1985) evaluated

the performance of 26 family practitioners in the care of clients with rheumatic disease. Covert SPs were utilised in this study and 88 per cent of the doctors were unaware that the patient was a simulation. A similar study was performed by O'Hagan, Davies, & Pears (1986) using 13 general practice teachers, 10 family practice registrars and 10 recent graduates. A covert male and female SP, portraying a migraine headache, were evaluated by these practitioners. Seventy-three percent of the doctors did not detect the case, 24 per cent were suspicious and only one doctor detected the case.

Norman et al. (1985c) evaluated the performance of three groups of family physicians using seven different covert simulated patients. The first group were doctors who participated in the development of some of the seven SP cases but were exposed to a case that they did not develop. The detection rate was 18 percent. The second group was a sample of physicians who volunteered to participate in the study and who were not involved in the development of the case. The detection rate was 21 per cent. The final group were doctors who participated in the development of the seven cases and then were later exposed to a case that they developed. The detection rate was 39 per cent; explained largely by case development recall. (Woodward, McConvey, Neufeld, Norman, & Walsh, 1985) also measured physician performance by having covert SPs gain access to the physicians' medical practices. Thirteen per cent of patients were detected and 13 per cent were suspected. False positives, where real patients are assumed to be SPs, were also noted by doctors but the percentage of this occurrence was not reported in the study.

In a comprehensive study conducted by Rethans, Drop, Sturmans, and van der Vleuten, (1991), 39 medical practitioners agreed to have their performance monitored using four covert SP cases. Twelve months later, these four SPs accessed the doctors' offices over the course of four months. The SPs were trained using standard methods and reliability of their presentation was monitored regularly. One hundred and fifty-six consultations took place and none of the SPs were detected. Tamblyn et al. (1992) also conducted a study using an approach similar to Rethans et al. above. Twenty-two physicians experienced four visits from covert SPs with histories of osteoarthritis. This particular study was unusual in that the patients

returned to the medical practice for a second visit. Unblinding of the SP occurred in 13.8 per cent of the cases largely due to problems with test results or atypical practice entry. Unblinding was not due to poor performance, and accuracy of patient presentation was 96 per cent (Tamblyn et al., 1992).

Students have also been asked to rate the fidelity of simulated patients. Gold et al. (1995) describes a SP program for medical students in geriatrics. Each student (n=108) spent twenty minutes with a SP and was asked afterwards to rate the fidelity of the simulation. Seventy-three per cent of the students found the experience to be real. Thirteen per cent were indifferent and 15 per cent found it artificial. Stillman et al. (1986) reports on a study where 1289 overt SP encounters took place over the course of a year with 336 medical residents. The SP encounters occurred throughout the year in an ambulatory clinic and students were advised to treat the SP as they would a regular client. Following the encounter, the residents completed a questionnaire asking them to evaluate the realism of the case. Eighty-nine per cent of the students felt that the simulation seemed real. Stillman et al. (1990) also provides the results of a study using fourth year medical students in four New England Medical Schools. Three hundred and eleven students saw between 15 - 19 SPs over the course of a day during an objective structured clinical examination. Post-experience questionnaires demonstrated that 87 per cent of the students felt the cases were realistic.

With respect to physiotherapy, Ladyshewsky and Gotjamanos (1997) had 73 physiotherapy students rate the realism of a SP encounter. One of four different SPs, each of whom presented with difficult social histories, were interviewed by each student. A five point Likert scale anchored as 1=not realistic to 5=very realistic was used to evaluate fidelity. The students rated the fidelity of the simulation as 4.31 on a 5 point scale.

In conclusion, the majority of studies using covert SPs attest to the fidelity of simulated patients. Three studies had perfect non-detection rates (Burri et al., 1976; Owen & Winkler, 1974; Rethans et al., 1991). Other studies had detection rates of less than 15 per cent (Gordon et al., 1988; McClure et al., 1985; Tamblyn et al., 1992; Woodward et al., 1985). Two studies by (O'Hagan et al., 1986) and (Norman et

al., 1985c) were less successful with detection rates greater than 15 per cent. Studies using overt SPs also provide evidence in favour of SP fidelity (Bowman et al., 1992; Gold et al., 1995; Stillman et al., 1990; Stillman et al., 1986). Champion (1989) notes, however, that students, because of their limited clinical experience, may be unable to accurately assess the fidelity of a simulated patient. The studies using more experienced clinicians, however, suggest that SPs are realistic.

Another aspect of the fidelity argument is whether practitioners interact with SPs in a realistic manner. The studies cited previously used detection rates as a measure of fidelity. Norman, Tugwell, and Feightner (1982) used a different approach to evaluate the validity of SP presentations. They evaluated the performance of medical residents using four real and four simulated patients. The real patients presented their own problem. Their same problems were also portrayed by a simulated patient. They found no statistically significant differences in the medical residents' performance across the two types of patients. Sanson-Fisher and Poole (1980) conducted a similar study only this time second year medical students' interpersonal communication skills were evaluated. Forty students conducted two 15 minute medical interviews and were advised that they would be seeing either genuine or simulated patients. While the training of the SP did not follow standardised methods, there were no statistically significant differences in the students' empathy scores between the real or simulated patients. The results of these two studies suggest that clinicians and students perform realistically in simulations. Because performance does not differ between real and SPs, the criterion validity of the simulated patient is supported.

3.3.2 Accuracy (reliability) of Simulated Patients

Accuracy or reliability of the SP's performance is another dimension that needs to be taken into consideration when employing this technology, particularly for research or evaluation of competency. Several reviews of the literature attest to the reliability of simulated patients (Champion, 1989; Colliver & Williams, 1993; Kinnersley & Pill, 1993; Norman et al., 1985a; Norman, Muzzin, Williams, & Swanson, 1985b; Swanson & Stillman, 1990; Vu & Barrows, 1994). Reliability or accuracy is important as variation in the representation of a finding makes a case a poor tool for

ongoing measurement studies. In an article by Tamblyn, Klass, Schnabl, and Kopelow (1991), they state that accuracy is present when there is reproducibility in the SP presentation. This is measured by recording the proportion of essential clinical features presented correctly in each SP encounter. Tamblyn et al. (1988, 1991) provide support for the accuracy of SPs in two studies. In the more recent study, the accuracy of SPs was evaluated using 839 student encounters. Twenty seven SP scenarios were depicted using 88 simulated patients across two medical schools. Average accuracy ranged between 90 - 94 per cent although some cases had scores less than 80 per cent. Accuracy was worse for physical examination findings and patient affect. One-third of the errors were systematic suggesting that these could be ameliorated with training.

A study by Vu, Steward, & Marcy (1987) also reports on the reliability of SP presentations. The question asked in this study was, what is the effect of a SP simulating the same problem throughout the day? Three third year medical students each conducted a standardised history taking and physical examination on three simulated patients. History findings were recorded on a checklist and the observed findings from the physical examination were recorded by the students. A high degree of accuracy was reported across all interactions. Incorrect history findings, as presented by the SPs, ranged from zero to four out of a possible 35 items. History findings were correct 94 - 97 per cent of the time in the first instance and 86 - 100 and 91 - 97 per cent of the time in the second and third instance respectively. Physical findings were also consistent except for one SP who erred in one finding in both the second and third instance. In response to these results, Vu and colleagues suggest that accuracy may have more to do with the quality and internal consistency of the simulator. Hence, SPs need to be observed regularly to ensure ongoing accuracy and should not portray more than 12 simulations in a day (25 minute simulations were used in this study). This amounts to a total of five hours of simulation per day.

These restraints on performance time to ensure reliability have been echoed by other investigators. Petrusa et al. (1987) noted that after 8 - 10 interviews, SPs have trouble remembering what they said to students and are less certain of their evaluations.

Stillman (1990) similarly stated that SPs can generally tolerate 10 - 15 encounters per day of 20 minutes duration each. It would appear from these authors that four to five hours of simulation per day is the maximum for any one SP. Ten students per day also appears to be a good cut off point to preserve the SP's evaluation reliability across the students.

To conclude, the review of the literature presented thus far, suggests that patient simulations possess good fidelity (content validity) and can present their case scenarios in a reliable and accurate manner. This makes them excellent tools for the measurement of student and clinician competence and performance.

3.3.3 Advantages of Simulated Patients in Evaluation and Research

Traditional forms of measurement within the academic component of health science programs typically utilise multiple choice questions (MCQ) and written examinations. These may be appropriate for measuring knowledge but often these testing formats measure data recall and sample only a small percentage of competencies (Ainsworth et al., 1991; Barrows, Williams, & Moy, 1987; Nayer, 1995; Stillman & Swanson, 1987; Williams et al., 1987). The MCQ format is also a passive test and frequently cues the student regarding the correct answer (Case & Swanson, 1993; Haydon et al., 1994; Miller, 1990; Nayer, 1995; Newble, Baxter, & Elmslie, 1979; Vu & Barrows, 1994). Clearly these methods of assessment have little value when trying to gain inferences about a student's competence or performance.

The patient management problem (PMP) is another method that has also been used to evaluate clinical competence. This method of evaluation was first developed in the 1960s in an effort to create more realistic evaluation of problem solving skills (Swanson, Norman, & Linn, 1995). This is a written or computer driven test which uses a clinical vignette and a series of staged questions that are elicited depending upon the actions taken by the student. This method has been criticised by Miller (1990), Page and Bordage (1995) and van der Vleuten and Newble (1995) who state it is difficult to find consensus among experts regarding the appropriate interventions and pathways for the test. Cueing is also a problem in this test and it does not discriminate candidates any better than more traditional MCQ formats (Page &

Bordage, 1995). Multiple choice questions, written tests and the PMP also do not involve interaction with a human subject, thus making their external validity with respect to clinical competence questionable.

Assessment of clinical performance using the clinical setting as the testing site also has low reliability because of the uncontrolled nature of the environment, the lack of standardisation of observers and clients, and the episodic and limited observation of the student (Ainsworth et al., 1991; Barrows et al., 1987; Helfer et al., 1975; Jain et al., 1997; McKnight et al., 1987; Miller, 1990; Stillman & Swanson, 1987; Vu & Barrows, 1994; Williams et al., 1987). The extent of episodic observation, for example, was evaluated by Gold et al. (1995). He surveyed a class of 119 medical students on the amount of direct observation they received on their clinical skills while in hospital practice. He achieved a 91 per cent response rate and found that nearly two-thirds of the class had very minimal or no direct evaluation of their skill by hospital based faculty. This rather shocking statistic may be an overestimate, however, because students: may be observed directly or indirectly without them knowing; may underestimate the amount of supervision they receive; or may have their own specific and varying definitions of what is considered to be supervision. One can only speculate whether this same situation exists in physiotherapy.

Conventional oral examinations, another form of evaluation used in the educational setting, involve an examiner asking a student questions about a specific diagnostic problem. This method of evaluation has been used for hundreds of years, particularly in British Commonwealth countries where the examination is seen as a rite of passage into the medical profession (Swanson et al., 1995). The conventional oral examination, however, does not measure the effectiveness of bedside manner or the students' ability to elicit an accurate history or perform a physical examination (Barrows & Abrahamson, 1964; Newble, Hoare, & Elmslie, 1981). Nayer (1995), in a review of the literature, also notes that the oral examination has low reliability due to inconsistency of the examiner, fluctuations in candidate anxiety, verbal fluency capabilities of the examinee and the use of nonstandardised questions. The predictive validity of the oral examination method is also questionable because only a few cases are used to evaluate the student's knowledge (Nayer, 1995). The poor

psychometric qualities of the oral examination resulted in their elimination from United States medical licensing examinations in 1965.

In light of the weaknesses associated with all of these evaluation methods, Williams et al. (1987) describe three main characteristics for the objective testing of clinical performance. First, tasks and instructions should be standardised so each student takes the same test. Second, the critical aspects of performance and the acceptable range of responses should be established in advance by the test author alone or through consensus of the test author and experts. Third, all scoring is performed in a standardised manner by the same examiner without knowledge of the examinee's prior performance. The use of the SP simulation in evaluation, unlike the other methods of evaluation previously described, may be more appropriate as a research and assessment tool as it meets many of the characteristics described by Williams et al. (1987). The patient simulation can be designed at a level suitable for the examination and level of the students. The use of SPs also allows for the environment to be controlled and minimises ethical and safety considerations. These advantages have all been cited by numerous educators and researchers (Ananthkrishnan, 1993; Barrows, 1993; Edwards et al., 1995; Edwards & Martin, 1989; Hannay, 1980; McDowell, Nardini, Negley, & White, 1984; McKnight et al., 1987; Nayer, 1993; Owen & Winkler, 1974; van der Vleuten & Swanson, 1990).

Tamblyn et al. (1988) and Helfer et al. (1975) state that the SP offers the advantage of permitting experimental control over the type of case(s) selected and the contribution of differences in patient population mix to variation in student or practitioner performance. Further, the use of SPs in performance based testing more clearly represents clinical reasoning and competence because it requires active behaviour on the part of the examinees (Haydon et al., 1994; Owen & Winkler, 1974). Smit and Van Der Molden, (1996) and Vu and Barrows (1994) argue in favour of performance based assessment as students can be observed directly in standardised situations which resemble those they will encounter in work practice. Performance based testing also covers the entire skill of conducting an interview not just mastery of component skills. A quote by Resnick L and Resnick D (1991), cited in Smit and Van Der Molden (1996), illustrates the rationale for performance based

assessment, “assessment of component skills, instead of the complex of skills, fails to recognise that complicated skills and competencies owe their complexity not just to the number of components they engage but also to interactions among the components and the heuristics calling upon them”. (p.42)

Skills which can be evaluated or investigated using SP testing formats include: interviewing; examination; communication and interpersonal skill; psychosocial assessment; reasoning and inquiry strategies; use of time; diagnosis; and patient treatment and education (Barrows, 1971; Barrows, 1987; Barrows, 1993; Edwards et al., 1995; Elstein, Loupe, & Erdmann, 1971; Elstein, Shulman, & Sprafka, 1978; Jain et al., 1997; McAvoy, 1988; McDowell et al., 1984; Stratford et al., 1990; Thomas-Edding, 1987). Simulated patients can also be used as a research tool to evaluate particular teaching strategies as they offer the only means of control when trying to evaluate the impact of a specific teaching strategy on the actual performance of students with patients (Barrows, 1987).

Even though SPs appear to offer incredible opportunities for increasing the fidelity of clinical problem solving examinations, Swanson et al. (1995) note that they are still a simulation. They state that no matter how realistic a simulation may be, it is still artificial and examinees may not behave as they would in real life. Simulations must inevitably leave out some aspects of the real environment thus leading to the possibility of behavioural artifacts appearing in the examinee’s performance.

3.3.4 Training Procedures for the Simulated Patient

Barrows (1987) describes a comprehensive method for training simulated patients. His methods for training SPs are cited in numerous studies, so much so, that his training protocols appear to be the gold standard for investigations using simulated patients. A brief overview of his training principles follow.

Three people are generally involved in the training of a simulated patient. They are the simulator, the expert and the coach. The simulator is the person trained to portray the patient case and should be the same gender and approximately the same age as the actual patient. While lay people can portray a SP, Barrows notes that they should have acting ability. Several investigators advocate the use of actors or individuals

with acting experience (Cohen, Gromoff, & Swartz, 1992; Jenkins, 1991; McAvoy, 1988; Nayer, 1993; O'Hagan et al., 1986; Ostrow, 1980). Jenkins (1991), for example, feels that actors can portray patients accurately and convincingly, can give accurate information and can bring a depth of human behaviour to the interaction that is real.

Because most simulations are based upon a real case, the expert is the clinician who was involved in the care and treatment of the real patient. They act as a resource to the simulator and the coach, who work together to try and bring the simulation to life. Stillman (1993) recommends that the trainer (coach) have a medical background. One hour training sessions are recommended and Barrows (1987) feels that a realistic simulation can be developed in three hours. The first hour is dedicated to orientation, the second to the practice of history and physical examination components, and the third to final review, fine tuning and dress rehearsal. Barrows states that the idea behind training is for the SP to learn all about a real patient so that it becomes their own problem, an approach that was developed by Stanislavsky called method acting. Scripts are not advocated by Barrows as the SP is coached to incorporate aspects of their own life and experience into the SP role. Stillman (1990), however, states that scripts can be used and can be exactly the same as the real case or be composed from a composite of cases. Both Barrows (1987) and Stillman (1990) agree though, that medical jargon and information about the case should not be provided to the simulator so they appear more medically naive: like a real patient.

Barrows (1987) and Woodward et al. (1985) note that intelligent and motivated individuals appear to make the best simulated patients. Recruiters must also attempt to find out if the person interested in portraying the case has a hidden agenda (Barrows, 1987; Coggan et al., 1980; Stillman, 1993). For example, a member of the community who has had a bad medical experience may want to portray a SP in order to show students how they should treat consumers. This would affect the fidelity of a simulation. Several additional considerations for SP recruitment and training are provided in the literature, specifically, physical and emotional suitability, punctuality and availability, comfort level regarding the possibility of disrobing, being examined

and/or being asked personal or intimate questions (Champion, 1989; McClure et al., 1985; Stillman, 1993).

With respect to training time, there appears to be considerable variation from the three hour timeframe cited by (Barrows, 1987). Champion (1989) states that simulations that require physical and psychosocial aspects to be portrayed may take in excess of 6 hours to train. Gold et al. (1995) reported training times of 8 - 15 hours for gerontology simulations. In a review of 24 articles using SPs for teaching and evaluation purposes by Jolly (1982), only six studies report on the time taken to train the simulator. Training times ranged from a low of 1 - 3 hours to a maximum of 20 - 25 hours. It would appear from the wide range of training times reported in the literature that training time is a function of the coaching skill available, the acting skill of the SP and the complexity of the role the SP is required to portray.

At the completion or as part of the training process, clinicians experienced in the management of the specific problem should examine the patient (Champion, 1989; McAvoy, 1988; O'Hagan et al., 1986; Stillman, 1990; Stillman et al., 1980). Where students are involved, both high and low ability level students should participate in the evaluation of the SP's performance. Inconsistencies in the simulation can be checked at this time and a discussion held to improve the fidelity of the simulation. Videotape can also be used to improve the performance of the SP by illustrating concepts to the simulator (McAvoy, 1988).

In conclusion, this section of the chapter has provided a review of the literature on simulated patients. The advantages of using this educational technology in education, performance evaluation and research have been outlined. A section of this chapter focussed on the fidelity or content validity and reliability of patient simulations. From this review it appears that SPs can simulate patient cases with a high degree of reality and accuracy. The next section of this chapter focusses on the specific methodological issues associated with the use of SPs in evaluation and research.

3.4 The Use of Simulated Patients in Research and Evaluation

This section of this chapter concentrates on the specific application of SPs in educational research and evaluation. The most common form of evaluation method

using SPs appears to be the objective structured clinical examination (OSCE). A large body of literature exists describing the OSCE and its psychometric properties. The issues that will be explored in this Chapter are as follows:

- Definitions of ‘competence’ and ‘performance’ and how these distinctions influence measurement;
- Factors influencing competence and performance outcomes, such as standard setting processes, first visit bias, case complexity and testing time;
- The use of single or multiple SPs to portray a single case and the influence this can have on competence or performance scores;
- Gender effects between the SP and examinees and the influence this has on competence and performance scores;
- Sequential testing of examinees over time (test security) and the influence this has on test scores;
- Case specificity and generalisability; and
- Student anxiety in SP based examinations.

3.5 The Objective Structured Clinical Examination (OSCE)

The OSCE was developed as a method for evaluating the clinical competence of medical students (Harden & Gleeson, 1979; Harden, Stevenson, Downie, & Wilson, 1975). It is now used widely in North American medical schools with over 80 per cent using the OSCE in their curriculum (Jain et al., 1997). Five physiotherapy schools in Canada also use the OSCE for the evaluation of practical skills (Nayer, 1995). The OSCE is also used as a component of the physiotherapy National licensing examination in Canada (Geddes & Crowe, 1998; Parker-Taillon, Cornwall, Cohen, & Rothman, 1992). Edwards and Martin (1989) and McKnight et al. (1987) also report on the use of an OSCE in occupational therapy and nursing student evaluation respectively.

The OSCE has also been used for determining medical licensure and entry to practice in a wide variety of clinical disciplines ranging from general medicine to psychiatry (Hodges, Regehr, Hanson, & McNaughton, 1998; Owen & Underwood, 1980; Retegui & Cornel-Avendano, 1999; Rothman & Cohen, 1995; Swartz & Colliver, 1996). It has also been used for the evaluation of final year and postgraduate medical students/residents (Barrows, 1993; Cohen et al., 1996; Colliver, Marcy, Travis, & Robbs, 1991b; Hilliard & Tallett, 1998; Jain et al., 1997; Skinner, Newton, & Curtis,

1997; Sloan et al., 1994; Stillman et al., 1991; Tamblyn, Klass, Schnabl, & Kopelow, 1991b). The formative and summative evaluation of clinical skills of undergraduate medical and allied health students has also been carried out using the OSCE (Harper et al., 1983; Jolly et al., 1996; Lloyd, Williams, Simonton, & Sherman, 1990; McKnight et al., 1987; Monaghan et al., 1997; Rosebraugh et al., 1996; Stratford et al., 1990). Modifications of the OSCE format have also been used for measuring the quality of care provided by medical practitioners in the workplace, for formative evaluation in geriatrics education and for the assessment of foreign trained medical graduates' interpersonal skills (Boulet et al., 1998; Fabiny et al., 1998; Norman et al., 1985b; O'Hagan, Davies, & Pears, 1986; Rethans, Sturmans, Drop, & van der Vleuten, 1991a; Tamblyn et al., 1992).

The OSCE is also purported to be a more valid and reliable form of assessment than traditional methods (Ananthkrishnan, 1993; Ferrell, 1995; Harden & Gleeson, 1979; Nayer, 1993; Robb & Rothman, 1985). The reason for the examination's higher validity is that all aspects of clinical competence are evaluated. The examination is also more reliable because two of the three sources of variation in the examination, namely patients and examiners, are controlled. The third source of variation, student performance, is therefore, evaluated much more accurately. While the OSCE proper, is not used in this thesis, many of the methodological principles associated with this evaluation format are utilised. Minimising the variation in the independent variables, eg. the situational context and the patient, is necessary if a valid measure of cohort performance is to be obtained. Given the principles inherent in the design of an OSCE station, its ability to control the amount of variation in the independent variables makes this method of evaluation attractive.

The OSCE is very similar to conventional 'bell-ringer' examinations commonly used in anatomy courses. For example, a student spends a fixed amount of time at a 'station' where he/she must answer a question related to the specimen on the table. When the time limit is reached, a bell is rung and the student rotates to the next station, as do all the other students in the examination. These same organisational principles apply to the objective structured clinical examination. A series of stations are established in which students are required to perform some function related to

clinical competence. Harden and Gleeson (1979) describe these competency skills as: history taking; physical examination; problem identification and differential diagnosis; interpretation of results; treatment planning; patient education; and interpersonal communication skill.

Two types of stations are commonly used in an OSCE (Harden & Gleeson, 1979). The first is a procedure station. This station assesses the students' history and physical examination skills. An examiner is present and scores the students' performance using a pre-established checklist. The second station is a question station. This station requires students' to answer questions from the previous station, for example, "interpret the findings you just elicited from the patient". When these two stations are linked they are referred to as a couplet. Harden and Gleeson (1979) recommend that an OSCE have approximately 20 stations in order to gain an adequate measure of an individual student's competence.

The implementation of the OSCE in research on clinical competence and performance has uncovered numerous issues related to the measurement of these two constructs. An overview of these issues follows.

3.5.1 Competence and Performance Constructs

Up until this point, the terms competence and performance have been used interchangeably. In reality, these two terms are quite distinct. Senior and Lloyd, cited in Rethans et al. (1991b) define performance as what a doctor does in daily practice as opposed to what they are capable of doing, which is competence. Barrows (1987) defines competence as the ability of subjects to demonstrate their performance when they know they are being tested. Thus, they are performing as well as they are capable. This is in contrast to performance which Barrows (1987) describes as everyday professional performance when subjects are not aware that they are being tested.

Rethans et al. (1990) state that researchers need to understand the distinction between competence and performance when designing experiments. Psychometrically, competence and performance are two different measurement points. In the examination setting, candidates generally behave to the best of their ability. This is

the competence setting. In the normal day-to-day situation, the doctor must deal with the context or performance setting which is rife with motivational and situational variables that do not exist in the examination setting. Rethans et al. (1990) found in a review of 18 studies poor consistency between investigators in their use of competence and performance definitions. Measurement instruments used to evaluate these constructs were also inconsistent. The authors state that for evaluation of performance, only unobtrusive measurement using covert SPs can be utilised.

The issue of whether one is measuring competence or performance has been investigated in a variety of studies. Norman et al. (1985b) conducted an investigation on medical practitioner quality of care using seven covert SP cases. The SPs saw both the practitioners who were involved in setting the performance criteria for the study as well as community based doctors who volunteered to participate in the study. Only 43 to 68 per cent of the criteria were performed by both groups of doctors. While the sample size of this study was limited it illustrates that performance in the practice setting may be different from what experts set as ideal care standards.

Rethans et al. (1991b) carried out a study to determine the differences between competence and performance using 36 general practitioners. Four SP cases were used. The doctors saw these SPs in their surgeries and then again five months later in an OSCE (different SPs portraying the same cases were used for the examination). National standards of care were selected as the performance criteria and broken down into obligatory, intermediate and superfluous categories by experts. The SPs rated the doctors' performance following each visit to the surgery using a criteria checklist. Scores from the stations in the examination were added together to yield the OSCE score. An efficiency score was calculated which was the total of obligatory actions divided by the total OSCE score. The efficiency score was then divided by the time score (duration of the intervention) to yield an efficiency by time score. Rethans et al. (1991b) found that the overall scores in the competency examination were consistently higher by nearly 50 per cent when compared to the performance scores. Efficiency and efficiency by time scores were also consistently higher, but in this case the higher scores occurred in the surgery (performance) setting. The authors

concluded that poor performance in practice does not mean poor competence as doctors may be performing more efficiently.

Kopelow et al. (1992) carried out a similar study. Twenty seven family practitioners were evaluated using 10 SP cases that were taken from a fourth year undergraduate medical student examination. Standard setting was done by one-third of this group and a group of internists who collectively reached at least 75 per cent agreement on the standards. The doctors were then broken down into three groups. The first group were the doctors who participated in the setting of the standards. They saw the SPs in their surgeries. The second group saw the SPs in their surgeries and then three to four months later saw these same patients in an objective structured clinical examination. The third group followed the same procedure as the second group but in the reverse order. Surgery visit performance was recorded by the SPs using a specifically designed checklist. Kopelow et al. (1992) found that all groups met only 30 - 60 per cent of standards. Overall performance was lower in the surgery setting for all 10 SP cases and was statistically significant at $p < .01$. Kopelow et al. (1992) explain the reasons for this poor performance. First, because only one episode of care was allowed in this study, measures of performance were possibly under-estimated. In actual practice, patients would return for follow up visits and additional essential information would be collected. This has implications for the establishment of standards in evaluation studies where only one episode of care may be utilised. Second, before standards can be set for a skills evaluation test, evaluators need to determine what actions a clinician might undertake in an examination setting.

The findings reported in these articles provide evidence for a distinction between performance and competence (Kopelow et al., 1992; Norman et al., 1985b; Rethans et al., 1991b; Rethans et al., 1990). The results of this review also suggest that the outcomes of this particular thesis will be focussing on the competence of physiotherapy students. The other noteworthy finding from these studies is that standard setting can be problematic. Clearly, candidates in these studies did not appear to reach the required competency or performance standards. The next section of this review, therefore, will focus on methods for developing acceptable standards in competency testing.

3.5.2 Standard Setting in Clinical Skills Evaluation

There are three specific issues related to standard setting in clinical skills evaluation. The first is whether the standards reflect actual practice. The second and third factors relate to case complexity and station duration. These two factors influence whether the candidate can actually achieve the standards in the context of the examination. In a review of the literature, Colliver and Williams (1993) note that the standards set for an examination are often higher than the average competence level exhibited by candidates. Quite often these criteria are set by experts and have no basis in actual practice for the average practitioner (Norman, Feightner, & Tugwell, 1983; Rethans et al., 1991a). Because of this expert bias, criteria should be set using a combination of evidence based practice and actual performance (Colliver & Williams, 1993) .

Standard setting procedures generally involve the use of experienced practitioners or experts. Croen and Moroff (1994) support the use of experts in establishing scoring protocols. They had eight blinded clinical educators review the written results of 32 students' post-encounter patient answers from an objective structured clinical examination. Each case was scored by 2 - 3 examiners using a four point scale. Raters were able to discriminate between high and low scoring candidates with good agreement. The intercase and inter-rater reliability co-efficients were 0.78 for communication and 0.84 for data gathering respectively. Croen and Moroff (1994) conclude that this ability to determine good versus bad performance validates the use of experts in standard setting exercises.

Norman et al. (1985b) used groupings of four family physicians and one specialist in order to determine performance criteria for a SP case. The family practitioners were asked to determine what critical information and actions were necessary for a family practitioner to achieve an acceptable performance in the context of a normal office visit. Specialists met with the investigators and stated their opinions on critical features in diagnosis and management. Rethans et al. (1991a, 1991b) selected four cases based upon a review of 24 nationally accepted and published Dutch primary care standards which were all common in general practice. These standards were broken down into obligatory, necessary, intermediate and superfluous activities. This method was chosen because the authors recognised the danger of asking experts to

set standards. Sloan et al. (1994) and Stillman (1993) also support this view. For example, in an OSCE for post-graduate residents, performance scores were quite low with a large proportion of students not meeting the 70 per cent standard (Sloan et al., 1994). Faculty bias in selecting examination items and establishing the ratio of critical to non-critical items was cited as one of the reasons for the poor success rate among post-graduate residents.

A different approach to standard setting is described by McClure et al. (1985) who assessed the clinical judgment of 26 family practitioners. Criteria checklists were developed by the investigators for 5 SP cases depicting rheumatic disease. Four consultants were then selected to review the checklists after having an opportunity to evaluate the simulated patient. Using the checklists as a guide, the SPs were then trained for 30 hours. Each SP was then assessed by four physicians who were aware that the case was a simulation but were unaware of the diagnosis. These encounters were videotaped and later viewed independently by three criterion judges who completed the criteria checklists. Inter-rater discrepancies amongst the judges were discussed as a group and a consensus rating was established for each item on the checklist. This consensus rating became the criterion for judging performance.

It is clear from these studies that standard setting for clinical skill evaluation is difficult. Ideally a combination of expert input and observations of actual practice need to be considered if valid criteria are to be established. Pilot testing is also important to ensure that the final standards are relevant. Quellmalz (1991) describes several principles that need to be taken into consideration when designing performance based assessments:

- Significance - do the criteria measure the depth of performance across domains;
- Fidelity - where possible, the tasks and conditions for the performance assessment should be as natural as possible;
- Expectations and quality levels - should be realistic;
- Criteria for judging the quality of performance - should be the same criteria that is applied when the task occurs under typical conditions;
- Developmental appropriateness - the benchmarks that are chosen should be appropriate to the developmental milestone of the candidate;

- Accessibility - the language used to specify the criteria should be clear and meaningful to users so everyone is clear about the meaning of terms.

Linn, Baker, and Dunbar (1991) also state that the design of performance assessments must also include: an analysis of the task; student familiarity with the problems; and ways students would attempt to solve them. This information is very much part of the development and standard setting process and needs to be considered in the development of a station.

3.5.3 First Visit Bias

Three specific variables can influence whether candidates are successful or not in meeting competency standards during testing. These variables are first visit bias, case complexity and station duration. Most competence or performance tests allow for only one interaction with the simulated patient. This may result in what Tamblyn et al. (1992) call first visit bias. First visit bias can have a negative impact on outcome scores if the case cannot be managed adequately in one interaction. The presence of first visit bias in performance testing has been documented in the literature by several authors (Kopelow et al., 1992; McClure et al., 1985; O'Hagan et al., 1986; Rethans et al., 1991a; Rethans et al., 1991b; Tamblyn et al., 1992). These studies all used covert SPs to evaluate the performance of medical practitioners. While the sample sizes used in these studies were small, making generalisation difficult, they do provide some useful information about first visit bias.

Tamblyn et al. (1992) used two SPs with osteoarthritis who visited a group of 22 medical practitioners on two separate occasions. First visit bias was measured by comparing the proportion of criteria met on the first visit with the combined proportion of criteria met for the first and second visits. A statistically significant difference was noted in both SP cases with the combined scores being 4.5 and 6.6 per cent higher than the first visit scores. In light of this result, Tamblyn et al. (1992) suggest that cases should be selected for performance testing that can be managed within one episode.

Physicians in a study by (O'Hagan et al., 1986) reported that the problem used for their study, a migraine case, was inappropriate for a standard ten minute consultation.

A longer patient encounter or multiple visits were needed in order to gain an accurate measure of the clinician's performance. Rethans et al. (1991a) also note that the performance levels recorded in their study may have been better if follow up visits had occurred. Clearly, case selection and the duration of practitioner-patient interaction are significant variables that can influence test results.

3.5.4 Station Duration

The other variable influencing the candidate's ability to meet the examination standards is station duration. There does not appear to be any formal standard for station duration in the literature. Instead, station duration is largely determined by the measurement objectives of the test. Most OSCEs, for example, use stations of approximately five to 20 minutes. This is because the OSCE only evaluates a component of a task. Multiple stations are used to sample the broad range of clinical skills in order to boost test reliability. Studies which have used SPs to evaluate the complete performance of students and medical practitioners generally use longer stations. Fewer cases are used than the OSCE because of the longer testing time needed to fully evaluate history taking, physical examination, communication, diagnosis and management skill.

Norman et al. (1989) conducted a study of medical practitioners using a variety of testing formats. Two of these formats included a twenty minute simulated office interaction with a SP and an objective structured clinical examination. The office interaction required the physicians to carry out a history, physical examination, and formulate a diagnosis and management plan. The simulated office interaction was better at discriminating between the three groups of physicians. Norman et al. (1989) state that the OSCE, therefore, may be a more useful discriminator at junior levels where basic skills are being evaluated. The extended SP interaction appeared to be more discriminating for the complex of skills.

Sloan, Bonnelly, & Schwartz (1993) carried out a study to see how student performance differed on the same case when evaluated using a 30 minute SP interaction versus an eight minute OSCE station. More deficits in student performance were noted in the eight minute station. They explain this by stating that

the OSCE may overestimate deficits because the examinees do not have enough time to implement all of the requisite tasks. Sloan et al. (1993) conclude that for a specific case, the longer SP interaction provides a more complete and reliable estimate of student performance. Bowman et al. (1992) provide further support for longer stations. They found a direct association between the duration of the interaction and the thoroughness of the assessment.

Shatzer et al. (1990) conducted an interesting study in which they exposed 15, second year medical students to an 11 station objective structured clinical examination. While only history taking skills were evaluated, five, ten and twenty minute stations were used. They found that students generally asked the same questions in the first five minutes. Variance among the candidates started to manifest itself when the 10 minute station was used and most students were most productive up to this point. The 20 minute station provided more information on efficiency as some students turned out to be more productive in the last ten minutes of the twenty minute station whereas others had been more productive in the first 10 minutes. Shatzer et al. (1990) conclude that a 10 minute station provided better estimates of both relative and absolute examinee ability.

Table 3.2 provides a summary of several studies and the station durations employed to assess practitioner skill. Both the OSCE station and practical skills evaluation formats are described. Only the time spent in the actual practitioner-patient encounter is recorded in Table 3.2. Time spent in post-encounter activities such as written tests, clinical reasoning questionnaires, or oral examination are not recorded as they can vary considerably depending upon the objectives of the study. In addition to these specific studies and their reported practitioner-patient encounter times, several reviews of the literature also provide some insight into the setting of station duration. Three large reviews of the literature on SP based examinations cite reports of successful SP tests and OSCEs with station lengths of 5 - 40 minutes (Swanson & Stillman, 1990; van der Vleuten & Swanson, 1990; Vu & Barrows, 1994).

Table 3.2: Summary of Station Duration in SP Evaluations

Reference	Station Duration (minutes)	Purpose	Task Structure	Sample Type
(Ainsworth et al., 1991; Connell et al., 1993; Colliver et al., 1991a; Harper et al., 1983; Haydon et al., 1994; Kaiser & Bauer, 1995; Lloyd et al., 1990; Sloan et al., 1993)	15 -90	Practical Skill Evaluation	Hx, Phys. Exam, Communication, Management	Medical students
(Barrows, Williams, & Moy, 1987; Colliver et al., 1991b; Gomez et al., 1997; Stillman et al., 1991; Stillman et al, 1990; Tamblyn et al. 1988)	10 - 20	OSCE	Focussed Hx & Phys. Exam, Communication, Management	Senior medical students
(Calhoun, Woolliscroft, & Ten Haken, 1987)	60	Practical Skill Evaluation	Hx, Phys. Exam., Communication, Management	2nd year internal medicine residents
(Edwards & Martin, 1989)	5	OSCE	Hx, Phys. Exam., Communication, Management	Occupational therapy students
(Jain et al., 1997)	15-30	OSCE	Hx, Phys. Exam., Communication, Management	Physiatry residents
(Ladyshewsky & Gotjamanos, 1997)	15	Practical Skill Evaluation	Focussed Hx	Physiotherapy students
(McDowell, Nardini, Negley, & White, 1984)	45	Practical Skill Evaluation	Hx and Phys. Exam.	Student nurses
(McKnight et al., 1987)	7	Practical Skill	Hx and Phys. Exam.	Level III and PG
(Nayer, 1995)	10 - 45	Practical Skill Evaluation	Hx and Phys. Exam., Treatment, Education and Communication	Physiotherapy students
(Petrusa et al., 1987; Rosebraugh et al., 1996)	4 - 8	OSCE	Hx. and Phys. Exam, Management	Undergraduate med. students
(Shatzer et al., 1990)	5, 10, 20	OSCE	History taking skills	2nd Yr med students
(Sloan et al., 1994)	5	OSCE	Focussed Hx or Phys. Exam.	3rd year med. students and surgery residents
(Stillman et al., 1986)	30	OSCE	Focussed Hx, Phys. Exam., Management	1st, 2nd & 3rd year medical residents
(Thomas-Edding, 1987)	30 - 45	Clinical Reasoning Study	History, Physical Examination	Expert physio & snr. physio students
(Williams et al., 1987)	20 - 30	OSCE	Focussed Hx, Phys. Exam, Management,	Final year med. students

(Hx = History, Phys. Exam. = Physical Examination)

It is clear from this literature that station duration can vary dramatically. van der Vleuten and Swanson (1990), therefore, state that station duration is dependent upon the aims of the study. Longer stations, for example, yield more measurement information but because of their length, introduce more measurement error and less reproducible scores. From a content validity perspective, however, van der Vleuten and Swanson (1990) and Miller (1990) state that evaluators must decide first what skills need to be evaluated. Once this is determined, task complexity and station duration can be determined to meet these objectives. Stillman, Regan, and Swanson (1987) also note that the time limit set for the evaluation must be adequate and appropriate to the students' level of training. A novice student, for example, would need more time to complete a task than a more tenured student or experienced clinician.

Few studies report on the use of SPs to evaluate undergraduate students' skill in implementing the entire patient care process. Connell et al. (1993) developed two SP cases in rheumatology to evaluate junior medical students' history, physical examination, clinical reasoning and decision making skills. Student-patient encounters of 45 minutes are reported in this study with up to an additional 25 minutes for post-encounter activities related to diagnostic reasoning and test interpretation. McDowell et al. (1984) used SPs to evaluate the patient care process in its entirety using nursing students. Forty-five minute student-patient interactions are reported. Haydon et al. (1994) evaluated 239 junior medical students using a SP case with hoarseness and cough. A 30 minute time limit was set for the student-patient encounter. Harper et al. (1983) used two cases portrayed by seven SPs to evaluate the clinical skills of undergraduate medical students. Students spent up to one hour on the case, but this included a student self-evaluation and discussing the case with two faculty observers.

From this review of the literature it appears that station duration is dependent upon testing objectives. For competency evaluations using the OSCE, numerous yet brief stations are employed in order to sample a broad range of skills. Comprehensive evaluation of the patient from history through to diagnosis and management,

however, require longer encounter times. Testing times greater than 30 minutes appear to be needed for this purpose.

3.5.5 Using Two or More Simulated Patients to Portray a Single Case

One of the issues that educational researchers have had to face in studies that attempt to evaluate competence or performance is whether to use one or several SPs to portray a single case. This issue usually surfaces when a large number of candidates need to be evaluated in a relatively short time frame. Having multiple people portraying a single case introduces variability in the testing situation since the portrayal of the case and the recording of scores by different SPs may not be the same across candidates.

The use of multiple SPs to portray a single case does not appear to be a problem in the OSCE provided the candidates are randomly assigned to the different SPs and there are enough stations in the examination (Colliver, Robbs, & Vu, 1991c; van der Vleuten & Swanson, 1990). At the single case level, however, scores may vary when using different simulated patients. Friedman et al. (1978), for example, state that similar problems may appear different to candidates when portrayed by multiple SPs because non-verbal cues would be different.

Colliver et al. (1991c) evaluated the scores of 210 students using 27 cases portrayed by multiple simulated patients. They measured the means of total, checklist and written scores for different SPs portraying the same case and whether the proportion of students failing differed between the simulated patients. For total and written scores, the effects of using multiple SPs was negligible. For checklist scores there were far more cases demonstrating statistically significant differences in mean scores across the simulated patients. Seven of the 23 cases had statistically significant differences with means differing from 10 - 35 points. Three out of 15 cases also demonstrated different failure rates between SPs portraying the same case. They recommend that for a given case considered in isolation, checklists completed by multiple SPs must be interpreted with caution.

Dawson-Saunders, Verhulst, Marcy, and Steward (1987) also studied the impact of multiple SPs portraying the same case in an examination and found that there is a

difference in student scores when different people portray the same SP case. If a written post-encounter score accompanies the checklist score then the total score difference appears to be lessened. They also state that the scores describing the student behaviours of patient education, presentation of management options and patient satisfaction ratings are more highly affected when multiple SPs are used. Focussed SP training and checklists that have minimal jargon are a couple of ways one can reduce this variability across simulated patients (Dawson-Saunders et al., 1987; van der Vleuten & Swanson, 1990).

Tamblyn et al. (1991b) looked at issues of reliability and bias by evaluating the outcomes of a shared OSCE at two different medical schools. A random sample of 537 out of 2560 encounters were videotaped and later rated by 44 SPs who presented the cases. Intra-rater reliability was better than inter-rater reliability within the same university and across the two universities. Tamblyn et al. (1991b) report that approximately seven to nine per cent of the reduced reliability was attributable to differences in item interpretation on the checklists when two or more SP raters were used. Poorer inter-rater reliability was also recorded when evaluations of patient teaching and communication had to be made. The conclusions that can be made from the research in this area suggest that where possible, a single SP be used to simulate a case. This serves to minimise variability. This recommendation is particularly relevant for this thesis project which uses a single case to evaluate differences in cohort performance.

3.5.6 Gender Bias in Evaluations using Simulated Patients

One important issue that requires some discussion is the influence of gender in performance evaluations. Colliver & Williams (1993) for example, note that female medical students generally do better in the area of interpersonal communication. Stillman, Regan, Swanson, and Haley (1992) evaluated differences in male and female post graduate medical residents' performance in an objective structured clinical examination. The only statistically significant difference occurred in year one residents, with females scoring higher on interview skills. Differences decreased as years of training progressed although women tended to have higher absolute scores for data gathering and interviewing and males for differential diagnosis and physical

findings. These two studies suggest that there may be some innate quality inherent in the students' gender that influence test scores. It is also possible, however, for the gender of the SP to have an influence on test scores. This concern is the focus of this particular section.

Colliver et al., (1991b, 1993) report on the influence of gender in SP examinations. Colliver et al. (1991b) evaluated the history and physical examination scores of six OSCEs, each of which had approximately 70 medical students and a total of 80 SP cases. There was a small, statistically significant student by SP gender interaction which the investigators attributed to a statistical artifact. Male students scored .02 points higher overall than females on male SP cases. Females scored .72 points higher overall than males on female SP cases. On a case by case basis, males had statistically significant better scores than females on three out of 36 male SP cases. Females had statistically significant better scores than males on three out of 40 female SP cases. Colliver et al. (1991b) concluded that there was no evidence of a gender interaction of any practical consequence.

Another similar study was conducted by Colliver et al. (1993). This study, however, looked at interpersonal communication. Four classes comprising approximately 70 medical students each were studied. There were no statistically significant overall interactions between student and SP gender for four of the five interpersonal communication skill categories being evaluated. Females, however, showed a statistically significant higher overall mean score for personal manner, although the difference was small. Statistically significant interactions were seen on a year by year basis but there were no consistent trends between males and females. Colliver and colleagues concluded that gender effects would be washed out in an OSCE but still encourage gender balance across stations where possible.

Other studies investigating the influence of gender on outcome scores during OSCE demonstrate similar outcomes (Furman, Colliver, and Galofre, 1993; Kuketich, 1992; Rutala, Witzke, Leko, & Fulginiti, 1991a; Gomez et al., 1997). For the most part, there do not appear to be any noteworthy gender biases.

Another reason why differences in male and female performance may surface may be due to the gender of the station (Colliver et al., 1991b). It may be that male students identify more positively with cases that are uniquely male and female students with cases that are uniquely female. For example, a SP portraying prostatitis would possibly be of more concern to male students whereas a female SP portraying menstrual problems may be of more concern to female students. Colliver et al. (1991b) independently identified cases that would potentially have a gender issue associated with the case and found that performance scores for male and female students at these stations were similar.

It would appear from these studies that the gender of the SP is a negligible concern in performance based testing where a large sample of stations is being used to evaluate candidates. These two points have been confirmed in a review of the literature (Swartz & Colliver, 1996). In studies using only a few SP cases, the possibility of gender effects is more real. One safeguard that can be employed to minimise this effect appears to be the avoidance of stations that have a clear gender bias. Studies employing only a few SP cases should also, where possible, analyse results by gender to ensure that no gender effect has influenced the outcomes of the study.

3.5.7 Sequential Testing and Examination Security

Evaluation of large numbers of students often requires testing to be carried out over a series of days or even weeks. This introduces the possibility of students breaching test security by advising other students of the examination's content. This issue has been investigated by a few medical educators. For example, Battles, Carpenter, McIntire, and Wagner (1994) examined the stability of an OSCE's scores over five years. While there were statistically significant differences between different years of administration, these differences were small and of no practical significance.

Colliver et al. (1991a) evaluated the results of five OSCEs over the course of five years. Each OSCE had approximately 70 senior medical students with 13 - 18 stations in the examination. Each class of 70 was broken down into five groups. Each group required three days of testing time. An entire class of 70 required 15 days of evaluation time. Only the scores from two SP cases out of a total of 83 cases

for the five OSCEs demonstrated a statistically significant change. A statistically significant linear trend was seen for only one of these cases. Only the overall OSCE score for students in the fifth OSCE administration was statistically significant ($p \leq 0.01$) from the scores of the other four groups. Colliver et al. (1991a) note that this was likely due to chance given the power of the statistical procedures. For overall test scores, which included both the practical and written components of the evaluation, Colliver et al. (1991a) conclude that there was no evidence of serious violations of test security. When the written component of the OSCE was reviewed, similar results were obtained. The practical components, which were scored using checklists, showed more statistically significant changes. For 67 SP cases, 15 had statistically significant differences in the overall mean scores ($p \leq 0.01$). Five of these cases showed statistically significant linear trends. However, four of these trends were decreasing in nature and one was increasing, a pattern inconsistent with breaches of test security.

Rutala et al. (1991b) also looked at the issue of test security. The overall scores of six serial administrations of an OSCE over an eight week period were evaluated. Only the skills component of the scores were investigated. The university's honor code was used to discourage the students from discussing the examination's content. Further, students were warned that they could give out inappropriate information which would jeopardise the performance of other students. A similar approach was used by Petrusa et al. (1987). They told students that problems would change during the OSCE to discourage them from trying to learn the identity of problems.

In the study by Rutala et al. (1991b) each station in the OSCE was 18 minutes. If a student finished a station earlier, they could leave the examination area and wait in a central lobby before the next changeover. There was also a lunch break in the middle of the examination. These break periods provided students with an opportunity for cueing. Given the examination's procedural aspects, Rutala et al. (1991b) evaluated the overall and individual station scores and conducted a similar comparison for morning and afternoon administrations of the examination. None of the stations showed a statistically significant change in scores over the course of a day or across the two months of test administration. It is important to note that this was the first

OSCE experience for the students in this study, although they all had exposure to SP cases. Students with more OSCE experience may be more able to pass on critical information.

Williams et al. (1987) also looked at the issue of test security by evaluating the scores of five groups of students who experienced an 18 station OSCE over two weeks. Each group encompassed 14 senior medical students. Group means did not demonstrate statistically significant differences across the groups that included both practical and written stations. Further, no statistically significant linear trends were evident across the five groups. In another study by Williams, Lloyd, and Simonton (1990), the nature of student communication about an OSCE was examined. They evaluated the OSCE scores of 156 medical students with no prior experience with simulated patients. A five case OSCE was used and the examination ran over 8.5 days with 10 students taking the test each half day. Following the OSCE, each student completed an anonymous questionnaire asking them for information about where they received information about the test. Seventy-six per cent of the students reported receiving information from other students. This information source was also cited as the most helpful. Orientation to the examination was the next most helpful source. The proportion of students receiving information about the examination also increased over the testing dates except for the last group. Interestingly, the scores of students did not increase over testing dates. In fact, scores of later test takers were generally lower than those who took it earlier. Williams et al. (1990) conclude that information received on the 'grapevine' does not automatically translate into better performance.

Williams et al. (1990) state that while students may have learned about the presenting diagnosis and patient presentation, they do not have an idea of the evaluation focus built into the case. Students in this study also did not have access to the evaluation checklists. Williams et al. (1990) also add that while the students do know the questions on the post-encounter probe, it is possible that knowledge of the diagnosis may lead to premature closure in resolving the case and result in lower data collection scores.

Reviews of the literature appear to conclude that breaches of test security do not appear to be a problem in OSCEs (Swartz & Colliver, 1996; Vu & Barrows, 1994). While the results of these investigations and reviews are encouraging, it is naive to think that students do not discuss their examinations or provide cues to one another. The limitations of statistical analysis, however, do not allow for investigation at this level (Colliver et al., 1991a; Rutala et al., 1991b).

Swartz, Colliver, Cohen, & Barrows (1994) conducted an interesting study whereby they studied the effects of deliberate, excessive violations of test security on a seven station SP examination using 140 fourth year medical students. The stations were of 30 minutes duration (20 minutes for the encounter and 10 for the post encounter probe). There were ten groups of students (14 per group) and students in group 8 were encouraged to divulge information about the test to students in group 9 through a variety of strategies outlined in the paper. The total differences in examination scores for all 10 groups were not statistically significant.

The issue of test security breaches in evaluations where only a few SP cases are employed has not been investigated to a great degree. Furman et al. (1997) repeated a seven case clinical practice examination ten times over the course of an academic year using 119 ambulatory care clerks in medicine. Clinical practice examinations are lengthy with the patient encounters being much longer than the typical objective structured clinical examination. For three of the cases, students received extensive feedback after the examination and were shown the checklists. Despite this information sharing, they still did not find any statistically significant differences for any of the seven questions over the 10 administrations of the test.

The results of the studies that have been reported in this section appear positive. It would seem that the OSCE and SP examinations may be relatively immune to test security violations caused by communication among examinees (Swartz & Colliver, 1996; Swartz et al., 1994).

3.5.8 Case Specificity and the Generalisability of Findings in Tests of Clinical Competence/Performance

Outcomes from research in clinical reasoning and clinical skills evaluation suggest that a single candidate's performance on one case is not a good predictor of performance on other cases. This is termed case or content specificity and has been described by numerous researchers (Elstein, 1995; Elstein, Shulman, & Sprafka, 1978; Elstein, Shulman, & Sprafka, 1990; Friedman & Mennin, 1991; Miller, 1990; Norman, Muzzin, Williams, & Swanson, 1985a; Petrusa et al., 1987; Rethans et al., 1991a; Rethans et al., 1991b; Rosebraugh et al., 1996; Swanson & Norcini, 1989; Swanson & Stillman, 1990; van der Vleuten, Norman, & De Graaf, 1991; Vu & Barrows, 1994). Case or content specificity was initially described by Elstein et al. (1978). In their study of clinical reasoning they found that expertise among medical practitioners was content specific and was limited to areas where the medical practitioner had substantial clinical experience. This phenomenon is not limited to medical practice and has been noted in the performance assessments of other disciplines (Linn et al., 1991; Vu & Barrows, 1994).

Content specificity, therefore, looks at the question of how does performance on one case correlate with performance on other cases. Correlations examining this question are generally low as evidenced in several reviews of the literature and in investigations designed to measure clinical performance (Norman et al., 1985a; Rethans et al., 1991a; Rethans et al., 1991b; Sloan et al., 1993; Stillman et al., 1986; Swanson & Norcini, 1989; van der Vleuten et al., 1991; Vu & Barrows, 1994). Given the case to case variability in examinee performance, researchers recommend that large numbers of encounters are needed (10-40) to make reliable and reproducible estimates of an individual's competence or performance (Harden & Gleeson, 1979; Norman et al., 1985a; Rethans et al., 1991a; Stratford et al., 1990; Swanson & Norcini, 1989; Swanson & Stillman, 1990).

Evaluation formats that use a few cases are evaluating group effects that are specific to the case(s) used in the study (Rethans et al., 1991a, 1991b). Even when one is evaluating performance differences among groups using a specific case, as is the nature of this thesis, one must ensure that one group is not more experienced than the other in the management of that specific case. This situation would result in one

group resolving the case using a process termed pattern recognition, whereas the other group may be using more in-depth backward reasoning skills (Friedman & Mennin, 1991). This scenario would obviously effect the interpretations of the performance assessment. Randomisation of the groups, sufficient sample size and a good understanding of the demographic backgrounds of the groups are needed to prevent the other factors influencing the results of the investigation.

In order to make generalisations about an individual's competence, and to compensate for the issue of content specificity, reliability coefficients for an OSCE in the order of .80 or more are generally recommended in order to have confidence in the examination (Colliver & Williams, 1993; Croen & Moroff, 1994; Ferrell, 1995; Nayer, 1993; Rothman & Cohen, 1995; van der Vleuten & Swanson, 1990). This generalisability coefficient means that an examinee's test score would be reproducible if a similar OSCE was to be administered to the same candidate. Factors that influence this reliability co-efficient are lack of inter-rater agreement, inconsistency in SP performance, inadequate numbers of stations on the examination, test length and method of score interpretation (Colliver & Williams, 1993; van der Vleuten & Swanson, 1990).

In a review of the literature, van der Vleuten and Swanson (1990) analysed the outcomes of 13 objective structured clinical examinations. They report that a minimum of 3 - 4 hours of testing time is needed to obtain minimally reproducible scores, regardless of station format. Similarly, Norman et al. (1985b) calculated intra-class correlation coefficients across the seven cases used in their study to evaluate physician performance. This coefficient was then used to determine the number of encounters necessary to achieve a reliability coefficient between 0.7 and 0.9. They report that a reliability coefficient of 0.7 could be achieved using any scoring method with seven to 36 encounters, but to achieve a reliability coefficient of 0.9, more than 30 encounters would be needed.

Norman et al. (1983) conducted an interesting study that looked at the generalisability of measures of clinical reasoning. They tested the concept of content specificity by evaluating the clinical competency scores of 30 medical practitioners, 10 specialists, 10 second year medical residents in internal medicine and 10 second

year undergraduate medical students. An OSCE with eight stations was developed. Two of the stations were exactly the same SP case, two stations had SPs presenting the same complaint, two stations had SPs presenting the same diagnosis, one SP presenting a different problem in the same specialty area, and one SP presenting a problem in a different specialty area. The outcomes of the study revealed little or no generalisability of scores between problems of highly similar or dissimilar content. Norman et al. (1983), therefore, question the view that clinical reasoning measures are case specific because two of the SP cases which represented the same problem were not found to have a high correlation. They concluded that the measures used to evaluate performance have an inherent unreliability and thus multiple problems are needed for individual assessment.

Stratford et al. (1990) evaluated the physical examination skills of 24 second year physical therapy students in a nine station musculoskeletal examination that focussed on soft tissue evaluation. They found that specific physical examination skills, when applied to a particular joint, had a moderate degree of generalisability to other joints. Contractile tissue tension testing being one example. However, other aspects of physical examination such as palpation or contractile pinching may not be generalisable to other joints. They advocate maximising the number of tasks and joint sites in a test to enhance the generalisability of an examination. Their findings support the concept of content specificity.

This summary of content specificity and the methods needed to ensure reproducibility of scores illustrates the rigour that is needed to evaluate an individual candidate's clinical practice ability. It appears that several hours of testing and the use of numerous cases may not be necessary if the research question is attempting to measure something other than an individual's clinical competence. Two examples are provided to illustrate this point. In the first example, one may ask the question, "how would a group of senior physiotherapy students reason through a selected case in neurology?" In this example, the study would be evaluating the cohorts' reasoning strategies in neurology and would make inferences using findings that are representative of the group. While the issue of case specificity would limit the findings to neurology and the case in question, the results would still contribute to

the clinical reasoning literature and would possibly reveal links to other studies in different specialty areas. The second example relates to the research question in this particular thesis, “how does clinical reasoning and skills performance differ when two cohorts of students evaluate the same case using an individualistic or reciprocal peer coaching approach?” In this example, the study is evaluating the impact of two different educational strategies on group performance. While the principle of case specificity would limit the findings of this study to the area of orthopaedics, the results would still contribute to our understanding of peer assisted learning in the health professions and hopefully illustrate links to similar studies in the literature.

The use of single or a few SPs to evaluate differences in group performance have been reported in the literature. Connell et al. (1993) evaluated the clinical competence of 19 junior medical clerks using two prolonged (up to 70 minutes) SP cases in rheumatology. The authors in this case were not trying to make generalisations about the competency of each student, but rather, used this experimental approach to study the cohorts’ reasoning and decision making skills in the context of an unfolding case. Day, Hewson, Kindy, and Van Kirk (1993) also used two SP cases to evaluate differences in strategic medical management among a group of first, second and third year internal medicine residents. Again, individual competency was not being assessed. Instead, differences in patient management by year of training was the focus of the study.

The outcome of this discussion on case specificity and generalisability illustrates that the use of the SP case has value both for investigations that attempt to evaluate individual candidate’s clinical competence and for measuring specific differences in group performance. The need to use a large number of shorter encounters or a few prolonged encounters is dependent upon the construct being measured and the research question under review. Researchers who use SP technology must, therefore, be aware of the various methods employed in studies of clinical practice and be cognisant of the phenomenon of content specificity when interpreting research findings.

3.5.9 Student Anxiety in Performance Based Testing

While student anxiety is not the focus of this thesis, it is a factor that can influence performance in SP based examinations. Sieber, O'Neil and Tobias (1977) cited in Champion (1989) describe anxiety as, 'the set of phenomenological, physiological and behavioural responses that accompany concern about possible failure in any testing or evaluative situation.' Anxiety by itself, however, can both facilitate or hinder learning. Yerks & Dodson (1908), for example, state that there is a curvilinear relationship between arousal and performance such that performance improves as arousal increases until a plateau is reached. Thereafter, continuing increases in arousal lead to decrements in performance.

Unfortunately, student anxiety in SP based examinations is not well-reported in the literature. Barrows (1971,1987) does state that one of the advantages of using SPs over real patients is that student anxiety may be reduced. These comments, however, are based largely upon the author's experience in using this educational technology. Champion (1989) conducted a major review of the SP and educational literature related to learner anxiety. She summarises that there are several possible reasons for a reduction in anxiety among students when interacting with simulated patients: there is more margin for error; second chances are given; and patients cannot be harmed. She notes, however, that students can also become quite anxious during SP interactions and that this will depend upon the instructional method, the student's past experience, the nature of the task and the experience of the actual SP encounter.

Champion (1989) actually provides evidence for a reduction in anxiety as a result of experiencing an interaction with a simulated patient. Two SPs portrayed clients with head injuries leading to physical and behavioural problems. One hundred and one students participated in this study and were divided into seven groups. Approximately 43 per cent of the 101 students watched videotaped excerpts of head injury patients with behavioural problems. The other 57 per cent used the SPs as part of their learning. This learning involved the SPs portraying their case in the tutorial using a time-in, time-out technique. This technique allowed the tutorial leader and students to stop the SP portrayal at anytime to clarify a point or ask a question. Students were then asked to think about treating people with closed head injuries

who present with behavioural problems. Students who experienced the SP role plays had statistically significant lower levels of anxiety ($P < 0.01$) following the interaction than those students who observed the videotape scenarios. Students in the SP group (72 per cent), however, had high/very high levels of anxiety during the tutorial in comparison to the students who observed patients on videotape (12 per cent). This study, while providing information on student anxiety and SPs, did not look at anxiety from the perspective of a direct (one to one) interaction with a simulated patient.

Ladyshevsky and Gotjamanos (1997) examined student anxiety during SP encounters. Seventy-three students conducted a 15 minute interview with one of four different overt simulated patients with difficult social histories. Following the interview, students were asked to evaluate their level of anxiety during the SP encounter using a 5 point Likert scale. The scale was anchored as 1=very anxious to 5=very comfortable. The mean rating was 3.07. Students also noted on the feedback sheets that their anxiety decreased as the encounter progressed.

Other studies provide more direct evidence or arguments related to student anxiety when personal (one to one) interactions with SPs take place. Soeters, Scherpbier, and van Lunsen (1987) tried to determine whether stress presented itself as an issue in an OSCE for medical students ($n=361$). They evaluated the failure rates of the examination which had 11 stations. Failure rates were higher than average for the first, fourth, sixth and last stations. The authors concluded that the results provided no evidence for the effect of stress on student performance. Gold et al. (1995) also report on student anxiety and the use of simulated patients. Sixty four medical students interviewed one of three geriatric SPs for twenty minutes. A post experience questionnaire was given to the students which asked students to rate their anxiety levels in working with the SP on a five point Likert scale. Anxiety was reported in 33 per cent of the students, whereas 17 per cent were indifferent and 30 per cent felt the SP was not anxiety provoking. It is important to note that the SP encounter was not formally graded in this study so student anxiety may be under-represented.

McKnight et al. (1987) also reports on anxiety levels in a non-compulsory OSCE for level III B.Sc.N. and post-graduate diploma nursing students. Seventy-seven students participated in the examination and 50 per cent found it to be more stressful than the usual evaluation system employed for the assessment of clinical skills, which was direct observation of clinical skills. Fifty per cent of the students, however, felt it was more objective. Edwards & Martin (1989) also report on the use of an OSCE in a Canadian occupational therapy program. They report that one of the disadvantages of the OSCE is that students remain anxious despite orientation to the format and pre-test practice sessions. They do not, however, provide any information as to how this was determined, nor do they compare it to their traditional oral examination process where students select a question and demonstrate the skill to an examiner.

It would seem that anxiety does manifest itself in SP based examinations and encounters. However, this has not been compared in any objective way to traditional methods of examination such as the oral examination or real patient encounters. Ytterberg et al. (1998) note that most studies demonstrate increased confidence in students when they gain hands-on patient experience with management responsibility. If this is the case, one could speculate that anxiety may decrease in students when dealing with real patients after having an opportunity to practice on a simulated patient. Ytterberg et al. (1998) provide some evidence for this claim. They found that the OSCE experience increased novice medical student confidence in their clinical skills to a statistically significant degree.

To effectively evaluate the impact of SPs on student anxiety and performance outcome, student anxiety should be measured prior to and after the SP encounter. This additional information can then be used to assist educators in the interpretation of examination results; for example, were anxiety levels extremely high and outcome scores very low?

This section of the chapter has provided an in-depth review of the uses of SP technology in research and evaluation. Much of the research stems from the use of SPs in objective structured clinical examinations. This examination format has yielded a large amount of information on the psychometrics of this testing format. Many investigators have also extended the principles of the OSCE and SP technology

to other studies. This is largely because of the opportunities the SP provides for standardising case experiences thereby eliminating much of the variability in studies of clinical performance. Numerous methodological issues, however, are associated with the use of the OSCE and simulated patients. Those of major significance, and methods for their minimisation, have been reviewed in this section. The next section of this chapter focuses on the instrumentation that can be used to measure competence/performance.

3.6 Methods and Principles for the Evaluation of Clinical Competence in Objectively Structured Clinical Examinations

This remaining section of this chapter will focus on the instrumentation used to evaluate clinical competence and performance. Typical methods used to collect information on examinee competence in objective structured clinical examinations are checklists, rating scales, open ended questions and multiple choice formats. This section, therefore, reviews a variety of issues related to the use of these measurement instruments. These issues are summarised as follows:

- Principles of station development in objective structured clinical examinations;
- Performance based testing and the use of subjective versus objectified measuring instruments;
- Rater reliability in clinical performance examinations, in particular, the use of simulated patients as evaluators;
- Methods (instruments) for the evaluation of history-taking, physical examination and interviewing skills in objectively structured clinical examinations; and
- Validity and reliability of SP based examinations

3.6.1 Station Development

The measurement of clinical competence generally requires evaluation of three distinct performance domains. These domains have been described by Bloom et al., (1956) and Shepard & Jensen (1999) as the cognitive, affective and psychomotor domains. The cognitive domain covers knowledge, comprehension, application, analysis, synthesis and evaluation. The affective domain covers principles of receiving and responding to phenomena as well as attributing worth to values, and organising them into a value system. The psychomotor domain relates to perception, the readiness to act, and the ability to guide a motor response from its early stages of

development through to more mechanistic and adaptive responses. The evaluation of these skills is possible in SP based examinations as all three aspects of Bloom's taxonomy are evaluated.

The development of a SP based test, as with any test, should be accompanied by a master plan outlining the goals and objectives of the educational test. This has been described as a table of specifications (Gronlund, 1981). Balla and Boyle (1994) describe several guidelines for planning performance assessments. They are: identify and describe the purpose(s) of the evaluation and the properties which will be assessed; identify and describe the information required and the types of instruments and techniques that will be necessary to collect information; specify details of how the assessment will be administered; describe the basis and methods of deriving the results and the approach that will be taken to report the results.

Barrows (1984) and Barrows, Williams, and Moy (1987) provide guidelines for the development of stations in clinical performance examinations. In general, they state that the important characteristics of the problem, performance objectives, observations that must be made to assess performance, and how the student will be scored, all need to be pre-established. Nayer (1993) notes that a station blueprint should describe the testing situation that will be presented to the examinee, the nature of the clinical problem to be evaluated, the level of student that will be assessed, the time allowed for the test, objectives to be tested, tasks to be performed, data to be observed, methods for observing and recording data, and the methods for scoring.

3.6.2 Subjective versus objective measurement instruments

A variety of different testing methods can be used to evaluate the cognitive, psychomotor and affective domains of candidates in clinical performance examinations. The most frequently used methods in OSCEs, however, are checklists and rating scales. Post-encounter evaluations in OSCEs also use a variety of different methods to evaluate the students' clinical reasoning skills, some of which are also used in written examination formats - eg. the multiple choice question. This format is often selected as it is seen to be more objective than more open-ended short answer formats (van der Vleuten, Norman, & De Graaf, 1991). All of these measuring

instruments, whether they be checklists, rating scales or post-encounter MCQ formats have varying degrees of objectivity.

The need to select the most objective method for evaluation would appear to be valid to most investigators. However, testing formats which are considered to be subjective, may in fact, yield similar information about candidates' performance with the same reliability as more objective tests. The question then arises, which measurement instrument does one select? This issue of objectivity in performance evaluation has been explored in detail by Norman, van der Vleuten, & De Graaff (1991) and van der Vleuten et al. (1991). These authors define objectivity as a goal of measurement which is free of subjective influences. However, subjective influences cannot be completely eliminated from any examination, therefore, another view of objectivity is suggested by van der Vleuten et al. (1991). They state that objectivity is a set of strategies designed to reduce measurement error. Hence, value free measurement is objectivity and the strategies used to achieve an apparently judgment free instrument is objectification.

Objectification is generally preferred by researchers because: it is seen as more valid and reliable; because rater biases are minimised; efficient because machines can often mark the outcomes; and transparent because the criteria and goals are made explicit at the start (Norman et al., 1991; van der Vleuten et al., 1991). In spite of this preference for objectification, van der Vleuten et al. (1991) argue that subjective testing methods are not necessarily inherently unreliable. In a selected review of the literature, for example, they found that rating scales were often as reliable as objectified checklists, with the latter having only slightly higher inter-rater reliability.

Checklist scores have also been shown to correlate well with more subjective global rating scales (Cohen, Colliver, Robb, & Swartz, 1997; Regehr, MacRae, Reznick, & Szalay, 1998; Van Luijk & Van der Vleuten, 1990). Van Luijk and Van der Vleuten (1990), for example, measured the inter-rater reliability of raters using checklists, global rating scales and specific item rating scales. Six hundred and eighty-three students were involved in the study and faculty raters were used to complete the checklists and rating scales. The inter-rater reliability of these three scoring mechanisms was 0.83, 0.72, and 0.72 respectively. The correlation between overall

test scores using the checklist method and the global rating method was also quite high (0.81). The investigators concluded that both tools are appropriate when norm referencing is being used, as both methods rank order students in a similar manner. The checklists, however, did provide more discriminating information as the scores were more dispersed and the standard deviation larger. Norman et al. (1989) also examined the reliability of checklist and global rating scales as a method for scoring clinical performance. Correlations using these two methods of scoring for individual station scores ranged from 0.64 to 0.92, and was 0.88 for the total test score. The findings from these two studies suggest that both methods may be suitable for performance evaluation.

Cohen et al. (1996) compared a 17 item checklist and a five item rating scale to evaluate the interpersonal and communication skills of 178 senior medical students in an objective structured clinical examination. Six SPs evaluated the students' performance using the checklist and rating scales. The correlation between the overall checklist and rating scale scores was 0.81. They concluded that both evaluation formats are psychometrically similar and evaluate the same things. This conclusion, however, can only be made when the content of the two instruments is similar because of the phenomenon of content specificity (Norman et al., 1991). The rating scales in the study by Cohen et al. (1996) were also seen to be more reliable than the checklist, with intercase reliabilities of 0.76 and 0.65 respectively. The higher reliability seen for the more subjective rating scales suggest that this format may be more appropriate for communication and interpersonal skill evaluation.

Harper et al. (1983) support the use of rating scales for communication and interpersonal skill evaluation. They used two SP cases to evaluate 92 students. A seventeen item checklist was used to evaluate interviewing, problem orientation and physical examination. A global rating score using a four point interval scale was also used. The overall global rating scores had better inter-observer correlations in comparison to the itemised checklist scores across the two cases. The global rating scores did not vary by case and were superior to the itemised observations of interviewing and problem orientation. For physical examination, however, the objectified checklists achieved a higher degree of reliability than the global

judgments. They concluded that for communication type skills, global ratings may be better because observer ratings tend to be influenced by their own perceptions of what is good versus poor communication skill. Physical examination skills, which are discrete, are more easily observed and captured by checklist formats.

The promising correlations between subjective and objectified methods suggest that either format may yield similar amounts of information. However, Norman et al. (1991) state that the instruments may not be measuring the same thing. For evaluation of basic clinical performance and isolated simple skills in undergraduate medical students, they suggest that both the global rating scale and checklist can be used because they essentially measure the same things. Checklists, however, are preferred as they provide more feedback, are more explicit, and define what is expected. If SPs are also being used to score the candidates' performance, then checklists are preferred because instrument completion requires less medical expertise. Rating scales, however, may be more appropriate for higher level students and post-graduates because efficiency and proficiency are generally more of a concern in the evaluation and can perhaps be better captured using global rating scales. This is certainly evident in a study by (Regehr et al., 1998). They evaluated 53 general surgery residents using an eight station OSCE. Each station was rated by two surgeons. Checklists, checklists with global rating scales, and global rating scales alone were all employed. They found that global rating scales scored by experts showed higher inter-station reliability, better construct validity and better concurrent validity than checklists Regehr et al. (1998).

In summary, both Norman et al. (1991) and van der Vleuten et al. (1991) feel that rather than pursuing objectivity and objectification at all costs in performance assessment, the use of rating scales or checklists should be based upon the educational level of the examinees, the skills that need to be tested, practicality, and acceptability. By approaching measurement in this way, it may be more appropriate to use what are considered more subjective measures in some circumstances versus more objectified measures.

3.6.3 Rater Reliability in Performance Based Examinations

The use of checklists and/or rating scales to evaluate clinical competence both require an observer to make a judgment about the examinee's performance. These judgments translate into a score which is seen to be a measure of the examinee's skill or competency. The use of expert clinicians and SPs as raters has been explored extensively in the literature. Given the interest in using a SP as a rater in this thesis, a review of their rating reliability follows. Adequate rater reliability on the part of the SP is important if they are to be used in the assessment of clinical competence (Champion, 1989).

In six comprehensive reviews of the literature on simulated patients and OSCEs, it is reported that trained SPs are able to rate the performance of candidates accurately (Colliver & Williams, 1993; Nayer, 1993; Norman, Muzzin, Williams, & Swanson, 1985; Swanson & Stillman, 1990; Vu & Barrows, 1994; Vu et al., 1992). Inter-rater reliability coefficients summarised in these studies range from 0.26 to levels beyond 0.95, depending upon the amount of training the raters receive and the standardisation of the measurement instruments/scoring procedures. This degree of accuracy is similar to the rating accuracy of medical practitioners who are involved in performance ratings, particularly for history and physical examination skills (Swanson & Stillman, 1990; Vu & Barrows, 1994). Furthermore, SP accuracy has been shown to be consistent over time (Vu & Barrows, 1994; Vu et al., 1992).

The researchers van der Vleuten and Swanson (1990) carried out a comprehensive review of the literature and specifically addresses the issue of the SP's inter-rater reliability using information from nine objective structured clinical examinations. Inter-rater reliability coefficients for history taking and physical examination skills generally fell into the range of 0.68 - 0.93 whereas education and communication skill checklists generally have lower correlations ranging from 0.5 - 0.77. One study is cited in this review, however, with poor inter-rater agreement ($r=0.42$) which was due to poor training and the use of videotape by raters to score checklists. Given the generally good reliability of SP raters, therefore, (van der Vleuten & Swanson, 1990) note that it is unnecessary to have more than one rater per station.

In spite of the relatively good inter-rater reliability coefficients reported in the literature, the inter-rater reliability of a specific SP rater still needs to be pre-determined prior to full scale testing to ensure confidence in the test results. Stillman et al. (1986) specifically evaluated the rater reliability of the SPs prior to and during the administration of an objective structured clinical examination. After the initial training in the cases, the SPs were each evaluated by four medical practitioners. A project staff member observed these interactions. After each interaction, the SP and the project staff member scored the checklist independently. The inter-rater reliability for history, physical examination and interviewing were, 0.7, 0.7 and 0.52 respectively. With additional training, the inter-rater reliability of the SPs during the OSCE was 0.82, 0.86 and 0.67 for history, physical examination and interviewing items respectively. Similar inter-rater reliability coefficients are reported in another study using fourth year medical students at four medical schools (Stillman et al., 1990).

Comparison of the SPs' ratings to medical practitioners (a criterion reference group) is one way of validating the reliability of the simulated patient. This approach has been used by a variety of investigators. However, it is important to note that considerable variation can also present itself among so-called expert raters. Elliot and Hickam (1987) evaluated the rating reliability of faculty physicians. Six patient instructors were each evaluated by four different students. These evaluations were directly observed by three faculty physicians. Both the patient instructors and physicians completed a 53 item checklist after each interaction. Faculty observers were unreliable in their assessment of physical examination skills 32 per cent of the time (poor agreement on 17/53 items). This was seen as quite high considering the physicians were familiar with the task, had limited environmental interruptions and used an objectified checklist. Newble, Hoare, and Elmslie (1981) also describe rather large variability in inter-rater reliability between experienced clinical examiners in a 90 minute objective structured clinical examination. Marker reliability for checklist items was poor, ranging from 0.25 to 0.77 despite rater training. Elliot and Hickam (1987) note that the areas of poor agreement among the faculty raters were generally in the area of palpation, something which is difficult to evaluate for an observer. De Champlain, Margolis, King, and Klass (1997) and Tamblyn, Klass, Schnabl, and

Kopelow (1991b) concur with this statement. This concern, however, did not appear in a study of physiotherapy students (Stratford et al., 1990). They had faculty examiners take turns acting as observer and SP in a practical examination for second year physical therapy students. The intra-class correlation coefficients of the eight faculty members who evaluated the 24 students was 0.82. Direct observation was not seen to be a hindrance for rating reliability in comparison to the faculty evaluator who acted as the SP, even though the SP had direct exposure to the tactile information of the test.

In spite of the possibility of low inter-rater reliability among expert evaluators, it is still useful to compare the rating reliability of SPs to this criterion reference group. This provides investigators with a measure of confidence in the rating ability of the simulated patient. The following describes a variety of inter-rater reliability coefficients between SP raters and a criterion reference group.

Rethans, Drop, Sturmans, and van der Vleuten (1991a) compared the inter-rater reliability of SPs against that of doctors. Prior to the SPs participating in the full study, three videotaped consultations were carried out for each simulated patient. After the consultation, the SP completed the checklist which evaluated the doctors' performance, and the three independent criterion judges scored the checklist by viewing the videotape. The reliability and consistency agreement scores all fell within the range of 0.9 - 1.0.

Bowman (1992) evaluated the inter-rater reliability of 13 SPs who were trained to present a sexually active client with specific health problems. Kappa coefficients were calculated comparing the SP scores to the scores derived by the investigators who scored the checklists using an audiotape recording of the interaction. The coefficients ranged from 0.33 - 0.61, with lower results for the interaction and education items and higher scores for the risk assessment and reduction advice items. The investigators note that the interpersonal and educational aspects of the interaction are less likely to have high correlations because of the innate subjectivity of these skills. Further, the investigators' scored the checklists using audiotapes which do not take into consideration any of the visual information that the SP would have experienced. This would contribute to the lower correlations seen in this study.

Calhoun, Woolliscroft, and Ten Haken (1987) conducted a study on the performance of second year internal medicine house officers. These house officers examined three real patients with stable chronic medical conditions. The patients completed checklists following the interaction and these scores were compared to the scores obtained by experts who observed the interactions on videotape. Statistically significant and positive correlations were seen between two of the three patients and the expert raters. Very few of the correlations, however, except for one, reached a correlation beyond 0.80. Part of the reason for not meeting this level of agreement may be due to the fact that real patients were used in this study. Barrows (1971,1987) and Jolly (1982) have stated that real patients vary so much within themselves that they are quite inappropriate as subjects for controlled evaluations. Secondly, the investigators note that the checklists were quite extensive. Excessively long formats have been shown to decrease reliability scores (Vu et al., 1992).

These studies, which describe the inter-rater reliability of SPs against that of expert practitioners, provide strong support for the SP as a rater. They suggest that the ratings of well trained SPs can parallel the ratings of clinical personnel. Another aspect of SP's reliability that must be considered, however, is that of intra-rater reliability. Intra-rater reliability is a measure of the SP's consistency in rating.

As part of a study on the competency of medical practitioners in rheumatology, McClure et al. (1985) measured the intra and inter-rater reliability of the simulated patient. After training, the SP completed four practice sessions with physicians who were unaware of the patient's identity as a standardised patient. These interactions were videotaped. The SP completed a checklist after each session and then two weeks later reviewed these same videotapes and completed the checklists again. This provided a measure of intra-rater reliability. Discrepancies were used to augment the SP training to the extent that all SPs who entered the clinical practices of the physicians in this study had demonstrable intra-rater reliability coefficients in the order of 0.85 or greater.

Stillman, Brown, Redfield, and Sabers (1977) also evaluated the intra-rater reliability of SP mothers engaged in an interview process with medical students. Reliability coefficients of 0.90 and 0.85 were reported for the two mothers who watched their

videotape performance twice on videotape, 2 weeks apart. Tamblyn et al. (1991b) also conducted an investigation of issues related to the intra-rater reliability of simulated patients. They compared the intra-rater reliability of 44 SPs by having them review their videotaped performances three months after the initial objective structured clinical examination. Original checklist scores were compared to these secondary scores. The average within-rater agreement in this study for individual checklist items was 82.2 percent. The average Kappa statistic, which corrects for chance agreement, was 0.52 for individual items and the average intra-class correlation coefficient for the total check list score was 0.37.

Tamblyn et al. (1991b) notes that these within-rater reliability estimates are likely under-estimated because the videotape review process is contextually different from the original scoring episode and does not provide the full spectrum of the experience. Further, SPs are likely to be more critical when viewing the videotape because they do not have to rely on recall to complete the checklist, nor do they have to perform the tasks. Simulated patients are also more likely to make errors of commission rather than omission in practical examinations by rewarding students for actions not done appropriately or by prompting the student into the correct course of action (Tamblyn, Klass, Schnabl, & Kopelow, 1991a; Tamblyn et al., 1991b). This obviously has an impact on scoring and can affect within and between rater reliability coefficients.

As is evident from this review on rater reliability, the use of videotape has been used extensively in studies designed to assess the inter and intra-rater reliability of the simulated patient (Calhoun et al., 1987; Connell, Sinacore, Schmid, Chang, & Perlman, 1993; MacRae et al., 1995; McClure et al., 1985; Mumford, Anderson, Cuerdon, & Scully, 1984; Rethans et al., 1991a; Tamblyn et al., 1991b; Vu et al., 1992). In most of these studies, the reliability coefficients have been moderate to good and the use of videotape has not appeared to present a big problem. Other studies, however, have noted that the use of videotape to record SP-candidate interactions can interfere with the measurement of rater-reliability (Connell et al., 1993; Jain et al., 1997; Liu et al., 1980; Stillman et al., 1991). Physical examination sometimes cannot be seen on the videotape (De Champlain et al., 1997). This would

negatively influence reliability scores. Tamblyn et al. (1991a) note that 12 per cent of physical examination findings in one particular study could not be assessed because of this visual impediment created by the use of videotape. To overcome this visual problem, double camera views have been recommended as one strategy (Connell et al., 1993; Stillman, 1993).

The results of these studies using videotape are clearly mixed. Some studies report quite good rater agreement using videotape as a method to capture external ratings. Others have had less success. It would appear that much is dependent upon the positioning of the videotape camera during recording, filming quality and the nature of the task itself. Where videotape is used to evaluate rater agreement, it may be necessary to accept lower levels of reliability because of the occasional visual problems that stem from using this technique. Otherwise, direct unobtrusive monitoring may be another option to overcome the problems associated with the poor visualisation that sometimes occurs with videotape.

Accurate measurements of rater reliability can also be influenced by other factors. The importance of rater training and the use of checklists that are clear, concise and free of jargon are also important factors for ensuring rater reliability (Champion, 1989; Dawson-Saunders, Verhulst, Marcy, & Steward, 1987; Ferrell, 1995; MacRae et al., 1995; Nayer, 1993; Tamblyn et al., 1991a; Tamblyn et al., 1991b; Vu et al., 1992). Stillman, Ruggill, and Sabers (1978) note that it took approximately 25 hours of training for a patient instructor to achieve an inter-rater reliability of 0.84 when checklist scores were compared to those of a faculty member scoring the interaction. Checklists which are too long also have a negative effect on rater reliability because it becomes harder for the rater to remember all facets of the interaction (Colliver & Williams, 1993; Ferrell, 1995; MacRae et al., 1995). Vu et al. (1992) recommends checklist lengths of 15 items as this provides for approximately 80 per cent accuracy.

The literature seems to suggest that inter-rater agreement tends to be higher for procedure oriented skills such as physical examination and history taking and lower for communication and patient education skills because the latter two categories involve more judgment and interpretation in completing the checklists (Bowman et al., 1992; Dawson-Saunders et al., 1987; Swanson & Norcini, 1989; Swanson &

Stillman, 1990; Tamblyn et al., 1991b). Communication skill (interviewing) scores, however, are more reproducible than data gathering skills because the former is less dependent on case specific knowledge (Swanson & Norcini, 1989).

The inherent accuracy of the SP can also influence rater reliability (Tamblyn et al., 1991b; Vu, Steward, & Marcy, 1987). Tamblyn et al. (1991b) explored the relationship between within-rater reliability and the accuracy with which SPs presented their problem during the examination. A strong and statistically significant linear relationship was present between accuracy of case presentation and within-rater reliability. Patients with case presentation accuracy scores of 80 per cent or lower, had a within-rater intraclass correlation coefficient of 0.14 whereas patients who were 100 per cent accurate had a within-rater reliability coefficient of 0.68. The actual characteristics of the SP such as consistency and accuracy may, therefore, have a strong influence on reliability scores.

Given the diversity of results that have been presented in this section, any study that makes use of raters, whether they be simulated patients or clinical examiners, must endeavor to establish the inter and intra-rater reliability of the evaluators prior to full scale examination. What is clear from this review, is that SPs can be used confidently as raters if they are carefully selected, trained appropriately and given checklists or rating scales that are clear and concise.

3.6.4 Methods (instruments) for the evaluation of history-taking, physical examination and communication skills

The measurement of history taking, physical examination and communication skills are commonly evaluated in tests of clinical competency. Norcini (1992) states that, “in complex clinical situations like those simulated by SPs, it is difficult to determine correctness with certainty...often several courses of action are correct (p.32).” Hence, there is an inherent difficulty in creating scoring tools in tests of clinical competence and performance (Swanson, Norman, & Linn, 1995). This next part of this chapter, therefore, examines a variety of specific instruments and scoring procedures which can be used to evaluate history taking, physical examination, communication (interviewing) skills, and clinical reasoning.

3.6.5 Checklists and Rating Scales

Checklists are commonly used to evaluate subjects in examinations employing simulated patients. Checklists need to be accurate and well designed to enhance discrimination between candidates by avoiding the potential to omit important items and including unimportant items (Miller, 1990; Stillman, 1993; Swanson & Norcini, 1989; Swanson & Stillman, 1990; van der Vleuten & Swanson, 1990). Omitting important items penalises examinees who are taking indicated actions and including unimportant items rewards examinees who are unjustifiably thorough. This is an important point but may not be appropriate for novices. Stillman (1993) notes that novices often go through more maneuvers in data collection than experts. As a result, checklists may need to be biased towards thoroughness rather than key features. The inclusion of key items only, however, makes it easier for the SP to remember what the candidate did during the encounter (Perkowski-Rogers, Solomon, Speer, & Ainsworth, 1992; Stillman, 1993; Swanson & Norcini, 1989). It would seem a balance between the two is needed for evaluation of novice competency.

Checklist items should generally describe one piece of information at a time, otherwise, rater reliability and the ability of the checklist to discriminate among candidates decreases (Colliver & Williams, 1993; Traub & Rowley, 1991; Vu et al., 1992). Checklists should also be free of jargon (Dawson-Saunders et al., 1987). Tamblyn et al. (1991b) evaluated 252 checklist items from 16 cases to determine what item characteristics led to poor inter-rater agreement. They categorised checklist items into three possible categories. First, unambiguous items generally contained one action per item. Second, explicitly ambiguous items had multiple actions per item. Third, implicitly ambiguous items were general and actually required the examiner to make a personal judgment. These latter two item types had the poorest rater agreement whereas simple documentation of an action produced the most agreement (Tamblyn et al., 1991b).

Other factors that influence instrument reliability and validity have been summarised by Williams et al. (1987). They state that in order to have confidence in the measures obtained by performance testing, the following procedures must be in place: tasks and instructions must be standardised; critical aspects of performance and the

acceptable range of responses should be established in advance by consensus of the case author and other experts; scoring should be carried out in a standardised manner without the knowledge of the examinee's prior performance; and examinees should only have their knowledge and skills to draw on when taking a test.

With respect to the actual scoring of checklists, a variety of different approaches have been used. Kaiser and Bauer (1995), for example, scored interactional skills on a 3 point scale (0=omitted to 3=outstanding). History taking and physical examination skills were scored dichotomously (0,1) as not asked/asked, or not done/done respectively. Similarly, (Sloan et al., 1994) used a 3 point scale (0=not done, 1=done poorly, 2=done well) on a checklist designed to evaluate medical student and surgical resident competency in surgical oncology. Rethans et al. (1991b) created three categories of scores for a performance checklist designed to evaluate the performance of general practitioners. They used obligatory (necessary), intermediate (not essential but not harmful) and superfluous categories to score their candidates differentially. An efficiency score was also derived by dividing the number of obligatory actions taken, by the total score. This efficiency score was then divided by the time taken to complete the interaction to yield an efficiency by time score.

Stillman et al. (1986, 1991) calculated unweighted history and physical examination checklist scores by calculating the percentage of checklist items obtained by examinees. They also obtained weighted scores by classifying each item on the checklist, using a six person study committee, as unimportant, indicated, important or essential with weights of zero, one, two and three respectively. These weighted scores were summed and divided by the maximal score and converted to a percentage. Critical scores were also calculated in the same way, the only difference being that weighted scores with values equal or greater to two were used in this analysis. They found that the weighted scores correlated perfectly with the unweighted scores. Critical scores correlated less well. Hence, they concluded that there was little value in weighting scores to derive a measure of the candidates' performance. Several researchers support this view noting that checklist item weighting does not have much impact on the reproducibility or validity of scoring as

long as the items themselves, are positively correlated (Page, Bordage, & Allen, 1995; van der Vleuten & Swanson, 1990; Vu & Barrows, 1994).

The length of checklists used in studies of clinical competence and performance also appear to vary. Vu et al. (1992) state that 15 items appears to be the optimum number for breadth of information and rating accuracy percentages of approximately 80 per cent. This was determined by evaluating a series of checklists with item numbers ranging from five to thirty. One factor repeated measures of analysis of variance were used to determine whether checklist length affected the SP's accuracy in recording. It was found that the effect of checklist length on SPs' accuracy of recording was statistically significant ($P=0.007$). The slope of the regression was -0.26 indicating that each time one item was added to the checklist, the SPs' percentage of accuracy in recording was reduced by 0.26. For every four items added to a checklist, therefore, rating accuracy decreased by one per cent.

More than 15 items may be needed on a checklist, however, depending upon the case in question, the measurement objectives of the evaluators, the duration of the encounter and the level of the candidate, eg. undergraduate versus post-graduate. This is evidenced by a variety of studies that used rather long checklist formats. Calhoun et al. (1987), for example, utilised three different checklists encompassing 83 to 116 items to evaluate internal medicine house officers' skills in evaluating real patients with chronic medical problems. Inter-rater reliability between the three patients' ratings and the two experts' ratings were 0.50, 0.56 and 0.28 respectively.

Elliot and Hickam (1987) used a 53 item checklist to evaluate the physical examination technique of 22 second year medical students. Statistically significant bias ($P<0.05$) between the responses of patient instructors and faculty members was observed for only five of the 53 physical examination items. Kaiser and Bauer (1995) used a 44 item checklist, scored by SPs to evaluate the interaction, data gathering and physical examination skills of medical students. This was compared to a self-evaluation checklist that the students completed on their own performance. Ninety-three per cent agreement between the overall SP and student ratings was reported.

The inter-rater reliability scores of these studies using longer checklists suggest that it may be possible to use longer checklist formats to evaluate history taking and physical examination skills and still obtain relatively good inter-rater agreement. What is probably important in these situations, however, is to ensure that the rater's have good training and that methods for enhancing rater reliability (described earlier in this chapter) are followed.

This section has focussed largely on checklists to measure history and physical examination skills. Many of the principles that have been discussed so far, however, also apply to rating scales. Nonetheless, a series of rating scales to evaluate interactional (interviewing and communication) skills have also been described in the literature. As noted earlier in this chapter, rating scales appear to be a more reliable and valid method for the evaluation of interactional skills. Norman et al. (1991) and Van Luijk and Van der Vleuten (1990) recommend the use of rating scales for interactional skill evaluation, as checklist formats, in their quest for objectification, often lead to the trivialisation of these skills.

Cohen et al. (1996) used a 5 point Likert rating scale to evaluate the interpersonal and communication skills of 178 medical students. The items evaluated clarity of communication, thoroughness of explanations, professional manner, personal manner and overall patient satisfaction. This rating scale was more reliable (intercase reliability of 0.76) than a 17 item checklist (intercase reliability of 0.65) used to evaluate these same communication attributes. In a similar study, Harper et al. (1983) also found that global rating scales, on average, had higher correlations than checklists for the evaluation of interviewing skills.

A Likert scale typically has a five to seven point 'degree-of-agreement' score which is anchored at two opposite extremes (Oppenheim, 1992). For example, a high scale score may mean favourable attitude whereas a low scale score may mean an unfavourable attitude. This would be scored as '5-strongly agree' and '1-strongly disagree' on a typical five point Likert scale. Oppenheim (1992) also notes that Likert scales generally have good reliability because of the greater range of answers permitted to respondents. Reliability co-efficients of 0.85 are often achieved. One criticism of the Likert scale is its lack of reproducibility as the same total score may

be obtained in many different ways (Oppenheim, 1992). Because of this, the pattern of responses is often more relevant than the total score. Scores in the middle also tend to be ambiguous making interpretation more difficult.

The Arizona Clinical Interviewing Rating Scale (ACIRS) has been used in several studies to evaluate the interviewing skills of medical practitioners (Calhoun et al., 1987; Day, Hewson, Kindy, & Van Kirk, 1993; Mumford et al., 1984; Stillman, 1980; Stillman et al., 1977; Stillman et al., 1983; Stillman et al., 1991). The ACIRS is an objective method of evaluating interviewing technique and encompasses 14 criteria which are applicable to any type of interview. It was developed by observing experienced clinicians considered to be excellent interviewers. Each of the 14 items is assessed using a five-point scale with descriptive anchoring statements describing excellent, average and poor performance in each area. The ACIRS covers the following six interviewing components: organisation; timeline; transitional statements; questioning skills; documentation of data; and rapport.

Tamblyn et al. (1991b) support the use of the ACIRS noting that it has good reliability because it is designed based upon what patients believe are necessary components of good communication. Further, raters merely indicate whether the behaviour is present or not. Stillman et al. (1977) report on several psychometric properties of the ACIRS instrument. First, inter-rater reliability and intra-rater reliability coefficients for the ACIRS are reported as 0.87 and 0.85-0.90 respectively. Secondly, the ACIRS possesses high construct validity as it could discriminate between medical students who had completed a paediatric clerkship from those who did not to a statistically significant degree ($p < 0.001$). Thirdly, the internal consistency of the instrument is also high, suggesting that the ACIRS does, in fact, measure one trait. The psychometric properties of this instrument, combined with its focus on generic interviewing technique, make it extremely attractive for evaluating the interactional skills of the sample populations in this thesis.

3.6.6 Methods (instruments) for the evaluation of post-encounter clinical reasoning

Analysis of candidates' clinical decision making skills are also necessary for a complete picture of clinical competence. The previous section looked largely at

measurement instruments designed to evaluate process skills. This section, therefore, focuses on evaluation methods that can be used in post-encounter situations to evaluate the clinical decision making skills of candidates.

A variety of different formats can be used to evaluate students' clinical decision making skills following an SP encounter. These range from very straightforward open ended questions to multiple choice formats (Connell et al., 1993; Harden & Gleeson, 1979). Newble, Baxter, and Elmslie (1979) compared objective (MCQ and true/false) testing formats to free response (open-ended) formats in examinations of clinical competence using medical students, interns and registrars. The objective testing format always resulted in higher scores in comparison to the free response format for similar and matched content examinations. Less competent candidates did much better in the objective testing format which suggests that this testing format over-estimates the candidate's clinical knowledge. Newble et al. (1979) state the free response format required spontaneous generation of knowledge, a more realistic requirement for clinical competence. The objective format emphasised recognition as the factor which led candidates to the correct answer. Being directed to the right answer simply by seeing it is called the cueing effect and is one reason why the MCQ format is considered to test a lower level of cognitive skill than the open ended format.

A number of elaborate methods for the measurement of clinical knowledge are also described in the literature. Many of these may be appropriate for the evaluation of clinical decision making. While many of these are used in written examination formats, they provide another option for evaluation of clinical decision making following a SP interaction. Solomon, Speer, Perkowski, & DiPette (1994), for example, describe an interesting format to evaluate the clinical reasoning skills of medical students in a standardised patient examination. They used an extended matching format which listed all the positive history findings that were key in leading to the diagnosis. Distractors were also included in the format. This method was selected to avoid some of the problems associated with coding open ended questions and the cueing problems associated with traditional MCQ formats. Students were required to identify the positive historical information which was key in leading

them to their diagnosis and to mark as many or as few of the selections as appropriate. Approximately 10 - 12 history data options were listed and fell into three categories: key findings that were in the SP's script that were necessary to elicit in order to obtain an accurate diagnosis; findings that were in the SP's script but were not relevant for determining the most likely diagnosis; and findings that were not in the SP's script. Key findings were weighted to yield a summary score equal to 100, not relevant findings were weighted by -10 and findings not in the script received weights of -25. One to four key findings were listed in each extended matching format.

The extended matching format is described in great detail as this format is used in the National Board of Medical Examiners and the United States Medical Licensing Examination (Case & Swanson, 1993). These question formats can be used to evaluate diagnostic skill, therapeutic skill or understanding of pathophysiology. A clinical vignette and lead in statement are generally used to provide the candidate with background and instructions for the question set. Candidates are required to select the single best answer or a specified set of answers to answer the question. These sets can range from 6 - 25 components. Case and Swanson (1993) note that the advantage of this examination format is that all the relevant options can be listed, rather than listing one correct option and having to invent three to four distractors which is what occurs in traditional multiple choice questions. Further, the MCQ format tends to favour low ability candidates because of the cueing associated with this testing format. The extended matching format, in contrast, forces examinees to construct an answer and locate this answer in the list of options that are provided. Case and Swanson (1993) argue that this leads to more discrimination in candidate evaluation, yields higher generalisability coefficients and results in more normal spread in the evaluation curve.

Later research expands upon the extended matching format, only this time it is referred to as the pick 'n' format (Ripkey, Case, & Swanson, 1996). This pick 'n' format was pilot tested on 9,782 students using 54 pick 'n' items in the second step of the United States Medical Licensing Examination (Ripkey et al., 1996). The design of the question is virtually the same as the extended matching format and

examinees must select several actions from a menu that lists approximately 10 - 15 items. The number of actions that must be selected is explicitly stated. Dichotomous scoring resulted in poorer discrimination among the candidates, therefore, the use of partial credit scoring systems is still supported (Ripkey et al., 1996; Rothman & Cohen, 1995). The pick 'n' format also resulted in higher discrimination among candidates than the 'one answer out of five MCQ' or 'patient management problem' format (Ripkey et al., 1996).

The extended matching format still appears to possess properties which cue the students. Woolliscroft, Swanson, and Case (1992), for example, compared the test scores of 164 junior medical students on a 23 item speeded test designed to evaluate pattern recognition. Each item described a patient condition and the students were required to select the correct diagnosis. A short answer test and an extended matching item test were used. Examinees had a higher mean score of 19.1 per cent on the extended matching format. This outcome was attributed, in part, to the extended matching format. The authors note that the short answer format provided richer information about the students' knowledge because of the lack of cueing.

Page & Bordage (1995) and Page et al. (1995) also describe an examination format for the evaluation of clinical decision making skills. It is called the key features examination. This research, which was carried out over six years, was funded by the Medical Council of Canada as part of its development for the Canadian Qualifying Examination in Medicine. They describe the concept of key features, which were originally described by Bordage and Page (1987) as the critical elements a candidate must consider in order to solve a problem. Without addressing these critical elements, the candidate is likely to be incorrect or unsuccessful in solving the problem. Like the extended matching format, key feature questions are useful for discriminating between candidates and minimise the problem of cueing which is prevalent in MCQ formats.

The key feature examination format is described in great detail (Page & Bordage, 1995; Page et al., 1995). First, the clinical situation and key features must be identified and written up, followed by pilot testing. Typical presentations should be used for entry-level candidates as atypical presentations lead to high failure rates.

Candidates and experts are asked to comment on the importance and difficulty of the items in order to ascertain face and content validity and to ensure that items listed on the questions are, in fact, key features. In their pilot studies they found that long menu selections led to more cueing. Shorter menus, averaging 15-20 items, were more reliable. For every correct key feature there should also be at least two other features to reduce guessing efforts. Partial credit scoring (number of correct responses from the menu or number of correct answers less incorrect answers) was also better than dichotomous scoring as the former led to higher discrimination and reliability.

Page and Bordage (1995) and Page et al. (1995) also state that specifying the number of selections that must be made to answer the questions (i.e., “2”) appears to be better than stating, “list up to a certain number (5)” as the latter situation forces students into a test taking strategy. For example, students may enter (2) correct and (3) incorrect answers in order to reach the maximum number of (5), when in fact (2) would have been appropriate. Page and colleagues also suggest that open ended questions be used for diagnostic and management decisions and be evaluated using marking keys with lists of acceptable synonyms. Questions about history and physical examination are better evaluated using the key feature menu format as the desired responses can be phrased more succinctly.

It would appear from this summary that a key features or extended matching item format would be a useful method for a post-encounter examination of the candidates’ clinical decision making skills in this thesis. Further, the use of partial credit scoring (unweighted) systems appear to be a more discriminating method for measuring differences in this domain. Open ended questions such as, “state your preferred diagnosis for this patient” or “list three important management strategies for this client” appear to be better for an evaluation of diagnostic and treatment skills as they provide richer information. These are harder to grade, however, as it is more difficult to score natural language (Woolliscroft et al., 1992).

3.7 Measures of Validity for Simulated Patient Examinations

3.7.1 Past Experience

In attempting to evaluate differences in clinical performance among individuals or groups, the influence of past experience can be a confounding variable. Background clinical experience needs to be controlled for in any experiment where you are trying to compare the competency of individuals or groups. If one of the individuals or groups has more clinical experience than the other, it may be this variable alone that leads to differences in individual or group performance.

The influence of past experience on outcomes scores has been investigated by several researchers. Blackwell and Callaway (1992), for example, evaluated the OSCE performance of medical students who had gained experience on certain sub-specialty wards in internal medicine to other medical students who did not have an experience on those wards. Twenty comparisons were made of the students' history, physical examination, diagnosis and management performance. Seven of the twenty comparisons were significantly higher for those students with relevant sub-specialty experience. Woolliscroft et al. (1992) had 164 junior medical students on internal medicine clerkships complete a pre- and post-test on 23 sets of items in which the students had to select a diagnosis. As students gained more experience in their rotations, improvements in the post-test scores became evident. Van Rossum, Briet, Bender, and Meinders (1990) also looked at the influence of prior patient experience on student performance in written tests and found that students who observed a specific case scored higher than those who did not observe the case.

3.7.2 Evaluating Validity in Performance Based Assessment

Moss (1992) describes some of the difficulties associated with the evaluation of validity in performance based assessments, namely, the substantial latitude that exists in interpreting performance and in the design of the tasks to be evaluated. Further, each response provided by a candidate is complex and reflects the integration of multiple skills and knowledge. Establishing measures of validity, particularly in the comparison of assessments, therefore, is problematic.

There is no one measure of validity, so one must demonstrate validity from a variety of different perspectives (Newble et al., 1981). Three types of validity generally need to be ascertained when developing a measurement instrument: content; criterion (concurrent and predictive); and construct (Gronlund, 1981; Newble et al., 1981). These three measures are inter-related and are part of a three part framework for evaluating validity as described by the 1985 standards of the American Psychological Association, the American Educational Research Association and the National Council on Measurement in Education (AERA, APA, & NCME, 1985).

Validity is concerned with designing appropriate and accurate tests of a target behaviour (Friedman et al., 1978). The 1974 standards of the American Psychological Association (APA, 1974) state that validity is, “the appropriateness of inferences from test scores or other forms of assessment,” (p.25). All three types of validity establish the extent to which differences in scores derived from the use of a test or technique will reflect true differences among individuals (or in the case of this thesis, groups) on the characteristics to be measured (Champion, 1989). In general, the use of OSCE examinations demonstrate good construct, criterion (modest) and content validity (Rothman & Cohen, 1995). This final section of the chapter looks predominately at construct and criterion validity from the perspective of SP based examinations.

3.7.3 Content Validity

The American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education describe content validity as “the degree to which a sample of items, tasks or questions on a test are representative of some defined universe or domain of content.”(p. 10-11) (AERA et al., 1985). Gronlund (1981) defines content validity as, “the extent to which a test measures a representative sample of the domain of tasks under consideration” (p.68).

Nayer (1993) and Swanson and Stillman (1990), in two reviews of the literature, state that SP examinations generally have high content validity because they are high fidelity simulations of real world situations. Gronlund (1981) also states that in criterion referenced tests, content validity is the most important determinant of a

test's validity. The content validity or fidelity of SPs was discussed in Chapter 2 and, therefore, will not be discussed in this section. Methods to ensure that both checklists and rating scales have good content validity were discussed earlier in this chapter.

3.7.4 Construct Validity

Moss (1992) states that

“the essential purpose of construct validity is to justify a particular interpretation of a test score by explaining the behaviour that the test score summarises. A strong program of construct validation requires an explicit conceptual framework, testable hypotheses, and multiple lines of relevant evidence to test the hypotheses” (p.233-234).

The American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education also provide a description of construct validity and note that:

“construct related evidence focuses primarily on the test score as a measure of the psychological characteristic of interest ... such characteristics are referred to as constructs because they are theoretical constructions about the nature of human behaviour ... the construct of interest for a particular test should be embedded in a conceptual framework. ... the conceptual framework specifies the meaning of the construct, distinguishes it from other constructs, and indicates how measures of the construct should relate to other variables.” (pp. 9-10) (AERA et al., 1985).

Gronlund (1981) describes several methods for evaluating construct validity. One of these methods is administering the test to known groups whose scores should differ as a result of training. This latter method has been used extensively in SP based examinations and is carried out by looking for differences in performance scores using candidates with differing levels of knowledge or training. For example, comparing the performance scores of junior to senior medical students, or one group of students who get specific training compared to those who do not receive any specific training.

Colliver and Williams (1993) conducted a comprehensive literature review on SP examinations and found that standardised tests using SPs have good construct validity with examination performance often being higher for groups with more training. Norman, Feightner, and Tugwell (1983), in a study designed to evaluate the concept of construct validity, evaluated 30 general practitioners, 10 specialists, 10 second year residents in internal medicine and 10 second year medical students in an 8 station objective structured clinical examination. All scores between the groups were highly significant for educational level, suggesting that the test was able to differentiate between levels of experience and training. Robb and Rothman (1985) also conducted a 20 station OSCE for year one, two and three residents and then followed this up with an oral examination four weeks later. They demonstrated construct validity of the OSCE examination with higher level residents scoring higher than lower level residents. Similar outcomes have been reported in other studies employing differing levels of students in objective structured clinical examinations (Newble et al., 1981; Petrusa et al., 1987; Stillman et al. 1986). In all cases students with more experience outperform those with less experience.

Not all SP based examinations report good construct validity. Day et al. (1993) evaluated the medical management of year 1, 2 and 3 internal medicine residents in an outpatient setting using two covert SP cases. Several measurement instruments were employed; a clinical reasoning rating scale was scored by the attending physicians on the residents' performance; a history and physical examination checklist as well as the ACIRS was scored by the SPs; and a medical management checklist was scored by the investigators who listened to an audiotape of the interactions. There was no appreciable variation in scores across all three years of residents for the checklists or ACIRS. The clinical reasoning skills rating scale also had uniformly high scores and only when the data from the two cases was combined was a statistically significant difference ($P < 0.05$) seen across years of residency. A variety of reasons for the lack of construct validity in this study are put forward. First, the lack of interpersonal skill variation may be due to inadequate teaching or poor reinforcement of these skills. The other possibility is that the SP cases did not have a strong enough emotional component so residents did not need to employ their skills. Day et al. (1993) also notes that the lack of any differentiation on the checklists

suggests that there may be no relationship between the amount of information gathered and experience. Given that these skills were not explicitly taught to the residents, they were all capable of implementing the skills to the same degree.

A study by Klass et al. (1990) also found little evidence for construct validity in an OSCE. They evaluated residents, medical students and unlicensed physicians from non-LBME accredited schools. While there was an increasing trend for higher scores with increased training level, the difference between medical students and residents was very small. No differences were seen across the groups on the interpersonal skill dimensions. They note that the lack of differences may be due to the nature of the measurement tool or scoring method.

As noted earlier, most studies using SP based examinations evaluate construct validity by measuring differences in performance scores when examinees with different levels of experience or training work through the SP case(s). Swanson and Stillman (1990) and van der Vleuten and Swanson (1990) disagree with using differences in scores by level of training/experience as the sole measure of construct validity. They feel that well constructed tests should differentiate between candidates simply on the basis of knowledge. A better measure of construct validity, therefore, would be to determine the degree to which differences should manifest themselves. van der Vleuten and Swanson (1990) state, however, if no difference is seen between groups with different knowledge or training, then the test is likely to be invalid.

This last comment may not be appropriate in all circumstances. Haydon et al. (1994) state that if certain skills or processes were never taught to the students then it is not possible to determine construct validity because the groups are inherently different. Jolly et al. (1996) provides evidence for this influence on construct validity. They correlated the OSCE scores of 152 medical students to 43 indices of the amount, nature, and quality of bedside, ward based or outpatient experiences. Only six indices correlated positively with the OSCE scores. Of particular note was whether students were able to examine patients on their own ($r=0.2$) and number of clinics attended ($r=0.18$). Students who performed a skill/task (even once) usually had higher mean scores for the relevant station. Seeing a skill or task performed usually did not confer much of an advantage in OSCE performance versus not seeing it.

It would appear from this review of research literature that SP based examinations do possess good construct validity when properly designed. Lack of construct validity can appear across groups with different experience levels if one particular group was never taught the specific skills being evaluated in the test. The amount of examinee experience within a group can also influence measures of construct validity. More experienced clinicians, for example, may not necessarily perform all of the tasks identified on a checklist but still arrive at the correct diagnosis. This is particularly true if the checklist is designed for a novice practitioner where thoroughness in technique is being evaluated.

3.7.5 Criterion Validity

Criterion validity applies when one wishes to infer from a test score an individual's (or groups') most probable standing on some other variable (Champion, 1989). This other variable is usually some other independent, valid and reliable assessment tool of the criteria being tested (Newble et al., 1981). This criterion related evidence demonstrates that test scores are systematically related to other outcome criteria (AERA et al., 1985). Gronlund (1981) defined criterion validity as, "the extent to which test performance is related to some other valued measure of performance". One of the difficulties associated with measuring criterion validity in clinical performance evaluations, however, is the lack of a gold standard related to clinical competence to which comparisons can be made (Gronlund, 1981; Nayer, 1993; Newble et al., 1981; Stillman et al., 1977; Swanson & Stillman, 1990; van der Vleuten & Swanson, 1990; Vu & Barrows, 1994).

There are two types of criterion validity that can be used to evaluate tests, one is concurrent validity and the other is predictive validity.

Concurrent Validity

Concurrent validity indicates the extent to which an individual's (or group's) standing on a criterion, compares to another test (Newble et al., 1981). If the results are comparable and the tests are measuring similar things, then concurrent validity may be claimed. In several reviews of the literature it is noted that correlations between SP examination scores and MCQ scores vary extensively (Nayer, 1993;

Swanson & Stillman, 1990; van der Vleuten & Swanson, 1990). These researchers also state that correlations between SP based examination scores and clinical performance ratings are generally in the low to moderate category. Swanson and Stillman (1990) and van der Vleuten and Swanson, (1990) state, however, that where positive correlations do exist, they do not necessarily mean that both tests are all evaluating the same things. The positive correlations only mean that the tests rank order candidates similarly. The content of the examinations must be assessed to determine whether they are measuring the same construct.

A few studies have been able to demonstrate concurrent validity. Norman et al. (1982) evaluated the validity of SP examinations by using four real and four simulated patients presenting the same clinical problems. No statistically significant differences were noted in the candidates' scores between these two patient groups and there was high overall agreement for the diagnoses for each patient group. Norman et al. (1982) argue that this is the only way to determine concurrent validity since real and SP cases encompass the full range of clinical performance. Sanson-Fisher and Poole (1980) were also able to demonstrate concurrent validity by comparing the performance of 40, second year medical students interviewing both simulated and real patients. Each student conducted two, 15 minute interviews with genuine, simulated or genuine/simulated patients. No statistically significant differences were noted in the students' empathy scores between the first or second interview suggesting that the students did not differ in their degree of empathy with either real or simulated patients; another measure of concurrent validity.

Robb and Rothman (1985) compared the scores of a 20 station OSCE to the scores obtained in an oral examination which was administered four weeks later to a group of first, second and third year residents. Two other testing formats were also compared to the OSCE and oral examination scores. These were the average results of the in-training evaluation reports and peer assessment ratings. Peer assessment ratings were derived by having the physicians assess each other's ability to perform as physicians. The OSCE correlated significantly with the in-training evaluation report and peer ratings, (0.54 and 0.72 respectively), whereas the oral examination

only correlated significantly with the peer rating (0.48). The results provide some support for criterion validity in SP based examinations.

Williams et al. (1987) correlated the OSCE scores of 70 senior medical students using 18 SP stations for history taking, physical examination, education, diagnosis and management to the scores of clerkship ratings and the National Board of Medical Examiners' (NBME) Examination - Parts I and II. The SP examination and clerkship rating scores were positively correlated (0.65) and much stronger than the correlations to the NBME examination scores (0.53 part I, and 0.51 part II). Petrusa et al. (1987) conducted a similar study and compared the OSCE scores of 204 medical students to their clerkship rating scores and the medicine component of part II of the NBME examinations. Moderate correlations of 0.46 and 0.43 were found to occur respectively.

Barrows et al. (1987) used an OSCE to evaluate the clinical competence of 72 senior medical students. The students' OSCE scores correlated well with their clerkship ratings ($r=0.65$) and eight students who were in the 90+ percentile for the OSCE were also among 12 students selected for the honors society of the program. These results suggest that the OSCE format and the processes used to evaluate students' competency were measuring the same constructs as those that were measured in clinical clerkship ratings.

Norman and Haynes, 1987 cited in Edwards and Martin (1989), provide some evidence for the criterion validity of the OSCE in physiotherapy. They correlated the scores of physiotherapy students from an OSCE to their scores obtained from clinical ratings. This correlation was 0.63 and provides a measure of criterion validity for SP examinations in physiotherapy.

Miller (1990) and Norman et al. (1991) note that, on average, most SP based tests have low correlations with other conventional tests (eg. MCQ), suggesting that they are measuring something different. Gomez et al. (1997), for example, demonstrated very low correlations between scores in a 10 station SP examination and academic scores in the medical/surgical curriculum. Correlations ranged between 0.15 to 0.24 in this study suggesting that the SP examination is measuring a different attribute of

clinical competence. Correlations with clinical ratings tend to be slightly higher and suggest that they are measuring the same critical qualities as those measured in a SP examination. This would appear to be the case from this review.

Predictive Validity

Predictive validity is the second component of criterion validity. Statements of predictive validity indicate the extent to which an individual's (or group's) future level on the criterion can be predicted from a prior test performance (Newble et al., 1981). Rapidly changing conditions may limit the usefulness of predictive validity in SP based examinations because examinees often receive further training or experience between the initial and retest situation (Champion, 1989). The logic of predictive validation also assumes that conditions existing at the start of the time sequence will continue unchanged until after the study is completed (APA, 1974). This is a rather difficult control to achieve in health science students whose clinical learning may evolve every time they are exposed to a new client. These factors make measurement of predictive validity difficult in the case of SP based examinations. Evaluation of later performance from earlier SP examination scores, therefore, is fraught with difficulty.

3.7.6 Internal Consistency

Gronlund (1981) states that the reliability of a test refers to a test's ability to reproduce the same result over time (consistency) and to be free from random errors of measurement. Internal consistency, is a form of test reliability which measures the extent to which the test items serve as a series of repeated measures of the same factors (Robb & Rothman, 1985). This may be achieved by splitting the test scores into two halves, for example, and comparing one half of the scores to the other half (Gronlund, 1981; Newble et al., 1981). If the scores correlate positively with one another, then the test is seen to be reliable (Newble et al., 1981). Measures of internal consistency, however, may be difficult to ascertain in a simulation as all students may not respond to the same items (Holzemer, Resnik, & Slichter, 1986). Nonetheless, internal consistency appears to be the best method for evaluating the reliability of the test scores in SP examinations. The test-retest approach is virtually impossible to carry out because of the time and expense involved in administering the initial

examination (Newble et al., 1981). Further, a significant amount of learning may occur during and after the initial SP-candidate encounter and would influence the test results if the procedure was repeated.

Several other factors can also influence reliability (Gronlund, 1981). Test length is one factor with longer tests yielding higher reliability because of the greater sampling that takes place. This also tends to produce a larger spread of scores which provides a greater measure of reliability in norm-referenced tests. The difficulty of the test is also an important factor. Tests that are too easy or too difficult restrict the spread of scores and the subsequent estimates of reliability. Criterion referenced tests, because they are not designed to emphasize differences among individuals, may present with a narrow spread of scores which will result in low correlational estimates of reliability.

In conclusion, this chapter has provided an in-depth look at the processes involved in establishing a simulated patient station using principles from objectively structured clinical examinations, in particular, issues related to the employment of simulated patients, the OSCE, and the design and development of checklists, rating scales and post-encounter tests. For the most part, simulated patients, the OSCE and associated instruments appear to yield scores that demonstrate adequate content, concurrent and construct validity for clinical competency measurement.

As this study is investigating the problem solving skills of physiotherapy students, an understanding of the nature of clinical reasoning is required. The next chapter of this literature review focuses on the clinical reasoning literature in medicine and allied health. A particular focus is the novice practitioner.

Chapter 4: Literature Review – Clinical Reasoning

4.1 Clinical Reasoning

Familiarity with the clinical reasoning (CR) literature is a necessity to understand how problems are solved in clinical practice. Given that this is one of the foci of this research, attention must be paid to the literature in this area. A review of the CR literature in medicine will be the major focus of this first section as this is the field of inquiry where most of the research has been conducted. Subsequent sections of this chapter will look at CR in other health disciplines followed by a review of CR in physiotherapy. Of particular interest is the reasoning of the novice professional practitioner.

4.1.1 Definitions

Numerous terms are used to describe the cognitive and procedural aspects of clinical reasoning: clinical decision making; clinical problem solving; diagnostic reasoning; decision analysis; clinical judgment; inquiry; heuristics, induction, prototypes, schemata, forward/inductive and backward/deductive reasoning, domain knowledge, hypothetico-deductive reasoning (Barrows & Feltovich, 1987; Barrows & Tamblyn, 1980; Higgs, 1990; Higgs, 1992; Newble, van der Vleuten, & Norman, 1995; Norman, 1985; Patel & Kaufman, 1995; Watts, 1985).

Barrows and Tamblyn (1980) prefer the term CR as it encompasses all the cognitive skills involved in patient evaluation and management. As noted earlier, most of the literature on CR comes from medicine, with the diagnostic skill of the clinician being the focus of study. This is quite different from CR in the allied health professions, whose focus on diagnosis is perhaps less salient. Carnevali (1995) provides a much broader definition and sees CR as a holistic process that leads to accurate and specific judgments, diagnoses and prognoses about a person's health status and situation. It enables the clinician to plan for rational, appropriate and individualised treatment.

Higgs and Jones (1995) also define CR very broadly, and describe it as the thinking and decision making processes associated with clinical practice. Clinical reasoning also occurs within, and is influenced by, several contexts: the personal context of the client; the context of the clinical setting; the personal and professional framework of the clinician; and the context of the health care system (Higgs, 1997; Higgs & Jones, 1995; Higgs & Titchen, 1995a; Jones, 1995). Further, CR is seen to encompass three core elements: knowledge, cognition and metacognition (Higgs & Jones, 1995). Cognition refers to the thinking skills of analysis, synthesis and evaluation of data, whereas metacognition is the awareness of thinking and the ability to assess one's knowledge base. The emphasis on the thinking and decision making processes, as opposed to the actions or steps involved in patient management, is another important distinction of the CR process (Higgs, 1992).

Jones (1992) defines CR from a physiotherapy perspective and sees it as a cognitive process used in the evaluation and management of a patient. This definition is more pertinent to physiotherapists as they must continually evaluate patients as part of their ongoing therapy, identify factors amenable to intervention, and effectively manage the problem (Jones, 1995).

The allied health professions have criticised the medical literature for ignoring the role of the patient in the CR process. This is seen in the following definition from Barker-Schwartz (1991). The author defines CR as:

“a complex intellectual process that surpasses logical thought and is depicted as a process that involves the therapist in a phenomenological approach to make sense of the patient's condition, and evokes the therapist's use of a caring perspective in establishing a collaborative relationship with the patient” (p. 1033). This phenomenological perspective to CR is more recent, and is reflected in the breadth of qualitative studies found in the allied health literature.

This brief overview illustrates the varying perspectives in defining clinical reasoning. While none of these definitions is incorrect, they reflect the scope of practice of the profession in question - i.e., diagnosis versus evaluation, treatment versus

management. Further, the method of study used to investigate CR has also influenced the development of definitions. For example, qualitative studies use more holistic and phenomenological descriptions of clinical reasoning. This is quite different from studies that have used more positivistic paradigms, whose definitions focus on the cognitive elements of decision making.

4.2 Clinical Reasoning in Medicine

For the most part, three theoretical models of reasoning are described in the literature: the hypothetico-deductive model of reasoning; pattern recognition models which relate to inductive or intuitive reasoning; and a third model which integrates aspects of knowledge with growing clinical experience (Higgs & Jones, 1995; Scott, 1996).

4.2.1 The Hypothetico-Deductive Reasoning Model

Elstein, Shulman, and Sprafka (1978) were some of the first researchers to analyse the CR of medical practitioners in a systematic manner. They, along with others, criticised much of the psychological research on problem solving as being too simplistic and lacking clinical fidelity for the complex process of diagnostic reasoning (Barrows & Bennett, 1972; Barrows & Feltovich, 1987; Neufeld, Norman, Feightner, & Barrows, 1981). Experiments in cognitive psychology, which studied the problem solving process in components, were seen as being too simplistic (Barrows & Bennett, 1972; Elstein et al., 1978). Experimental methods that reproduced the actual process of clinical decision making, using simulated patients, were seen as more valid ways of studying CR (Barrows & Feltovich, 1987).

In a significant work by Elstein et al. (1978), they used overt simulated patient (SP) cases to study the CR processes of 24 doctors. Both experts (as nominated by their peers - criterion group) and non-experts (not nominated by their peers - non-criterion group) were studied. The doctors were encouraged to think aloud during the patient encounters, as well as afterward, while reviewing their performance on videotape. The verbal reports of the doctors were subsequently analysed. From this work, Elstein et al. (1978) developed a model of CR which they termed the hypothetico-deductive (H-D) reasoning model. It involves four stages: cue acquisition, hypothesis

generation, cue interpretation and hypothesis evaluation. They found that the doctors generated a series of hypotheses which directed their inquiry strategies. The hypotheses could be general and become more specific and complex, or could be very specific from the outset, possibly in competition with others. The total number of hypotheses considered by the doctors at any one point in time varied from between four to seven. This model of reasoning has been very influential in guiding the research in clinical reasoning.

Barrows and Tamblyn (1980) provide an in-depth description of this model which is presented in Figure 4.1. Rather than being a four stage linear process, it is depicted as a cyclical process. Barrows and Feltovich (1987) and Barrows and Tamblyn (1980) also provide additional information on this model of reasoning. For example, they argue that hypotheses are generally based upon past experience with patient problems, and are processed in parallel with other hypotheses. Further, hypotheses can be in the form of a diagnosis or can represent anatomical, physiological, pathophysiological or aetiological explanations for problems. In addition, the concluding hypothesis is often generated early in the encounter with remaining questions used to build rapport, minimise reasoning errors and to inquire about potential management options.

Support for the H-D model is also seen in the following studies. Barrows et al. (1982) studied the CR of 37 general practitioners and internists using four SPs and methods similar to Elstein et al. (1978). Multiple hypothesis generation, early generation of the correct hypothesis and the preferential use of confirmatory evidence to support hypotheses was found in this study. Neufeld et al. (1981), in contrast, evaluated the CR skills of 35 medical students in a problem based curriculum. The results were contrasted with a similar study carried out on certified medical practitioners. Both of these studies used methods similar to those used by Elstein et al. (1978).

All groups in both studies entertained similar numbers of hypotheses and preferentially used confirmatory evidence in their reasoning. The first hypothesis appeared early in both groups. These studies revealed that the H-D model of

reasoning appeared to be the process used by both students and doctors. The only difference was that more experienced clinicians produced more detailed and specific hypotheses.

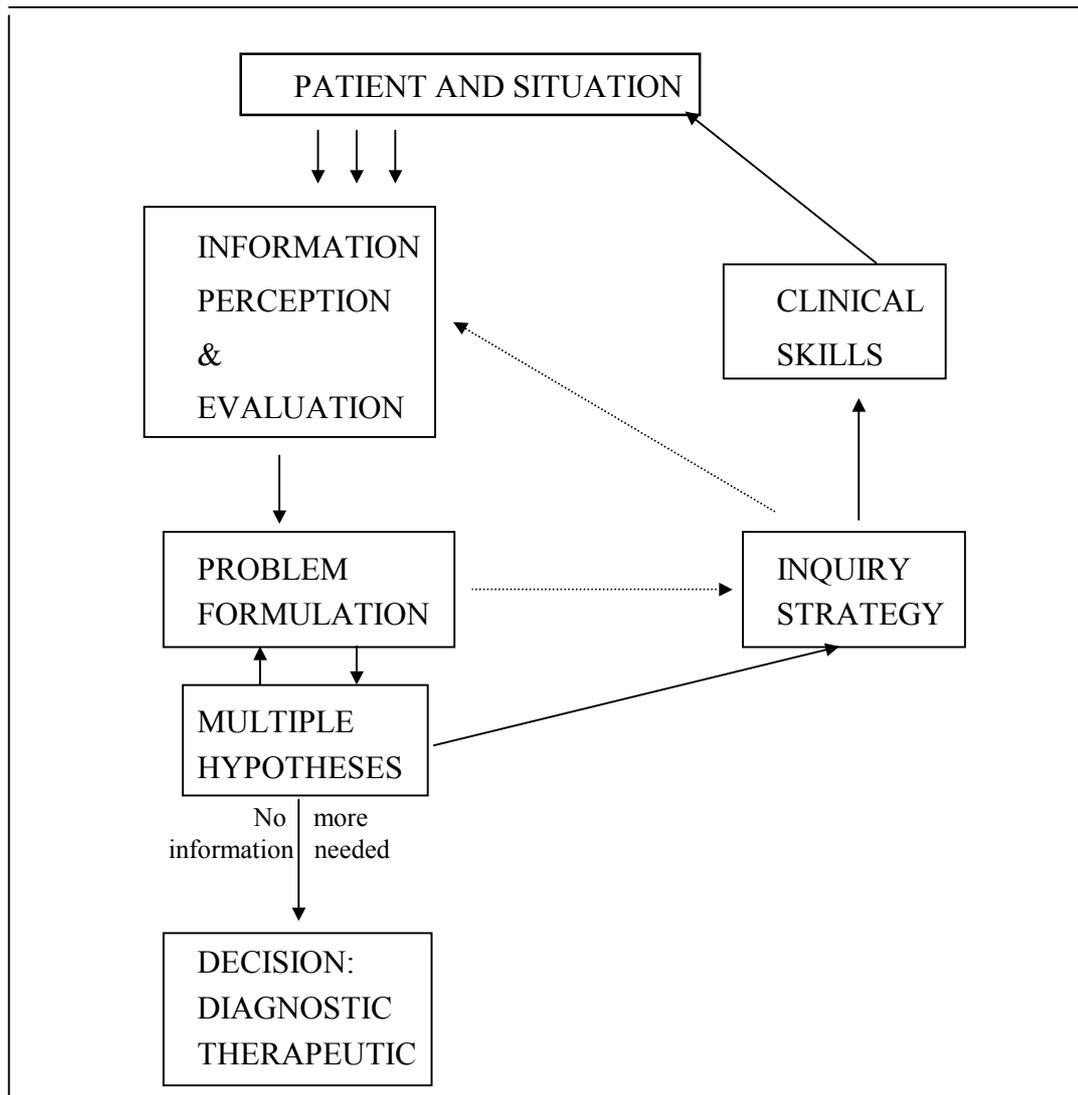


Figure 4.1: The Clinical Reasoning Process
 (Source: Barrows & Tamblyn, 1980)

In a study by Gale (1982), she argues that the H-D model of reasoning is characteristic of mature adult thinking and that an in-depth analysis of the cognitive strategies used in H-D reasoning was lacking. Bordage and Lemieux (1986) agreed with this argument and stated that the H-D model was macroscopic and did not delineate the cognitive strategies that were employed in each of Elstein's four stages. Gale (1982), therefore, used content analysis to study the verbal protocols of 66 medical practitioners each involved in the evaluation of a patient. These encounters were videotaped and played back to the doctors and stimulated recall was used to

identify aspects of diagnostic thinking. Fourteen categories of diagnostic thinking were identified which provide a much richer description of the thinking processes used in Elstein's four stage H-D model. These categories are described in Table 4.1.

Table 4.1: Categories of Types of Thinking in Clinical Problem Solving

Category 1a: Pre-Diagnostic Interpretation of Clinical Information: where the subject makes some active interpretation of the clinical information which would be insufficient as a diagnosis.

Category 1b: Diagnostic Interpretation of Clinical Information: where the subject makes some active interpretation of the clinical information which would be sufficient as a diagnosis.

Category 2: Judgment of need for further general or clarifying inquiry, not stemming from 1a or 1b thinking: Where the subject inquires further about the patient's symptoms, signs, etc. for clarification or where the subject seeks to clarify the patient's statement. N.B. not where the subject is seeking a particular piece of information based on his own expectations

Category 3: Expecting, searching for or planning to search for specific features (symptoms, signs, tests, etc.) of disease or treatment of disease: Where the subject shows expectation of certain clinical information or considers certain features of disease likely or possible in the patient, given the information already elicited.

Category 4a: Reinterpretation of clinical information, when no new information has been added: Where an array of clinical information which has already been interpreted in some way becomes amenable to a new (altered or additional) interpretation because of a change in the subject's own thinking and not because new information has been added to the array.

Category 4b: Reinterpretation of clinical information arising from the addition of new information: Where an array of clinical information which has already been interpreted in some way becomes amenable to a new (altered or additional) interpretation because of the addition of new information to the array.

Category 5a: Active confirmation of an interpretation: Where the subject feels that the selected interpretation is confirmed as an actual diagnosis.

Category 5b*: Active elimination of an interpretation: Where the subject eliminates an identified interpretation because of contrary evidence or positive lack of necessary evidence.

Category 5c: Postponement of either confirmation or elimination of a possible interpretation with or without stated differential likelihoods: Where an identified possible interpretation is neither confirmed nor eliminated by the subject but is left under postponed judgment.

Category 6a: Patient-determined interview structure: Where the course of the interview as directed by the subject is determined by, or follows on from, the flow of information as presented by the patient.

Category 6b: Subject-determined interview structure: Where the course of the interview is determined by the subject's requirement actively to test his interpretations of the clinical information.

Category 6c: Logically-determined interview structure: Where the subject conducts, or attempts to conduct, the interview according to a routine format as defined by the standard (taught) clinical history of any of its component parts.

Category 7a: Failure to make specific inquiry: Where the subject identifies, in retrospect, his own failure to make relevant, specific inquiry concerning the patient's problem, symptoms, signs, etc.

Category 7b: Failure to make general inquiry: Where the subject identifies, in retrospect, his own failure to make sufficient routine, general or screening inquiry.

Source: Gale (1982)

The H-D model of CR has since been expanded upon by other investigators. For example, the CR model of Barrows & Tamblyn (1980) has been expanded to include the concepts of cognition and metacognition (Higgs, 1990; Higgs, 1992; Higgs & Jones, 1995; Jones, 1992). This adapted model, along with a conceptual framework outlining the CR process in action, are provided in Figures 4.2 and 4.3 respectively. Again, these models represent the H-D model as more cyclical in nature, in comparison to the original linear model described by Elstein et al. (1978).

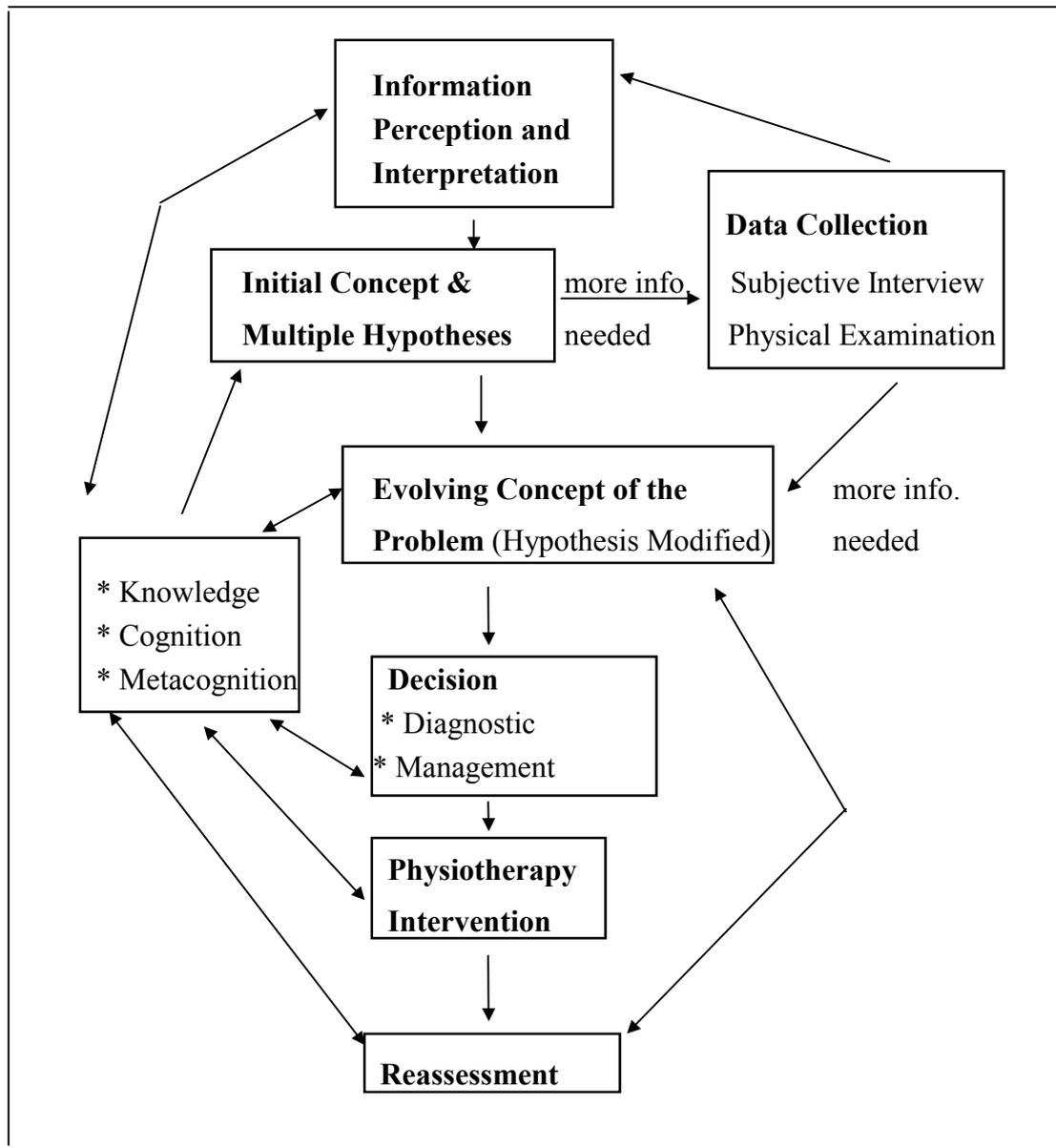


Figure 4.2: The Clinical Reasoning Process

Source: Higgs & Jones (1995; Jones, 1992) adapted from Barrows and Tamblyn (1980).

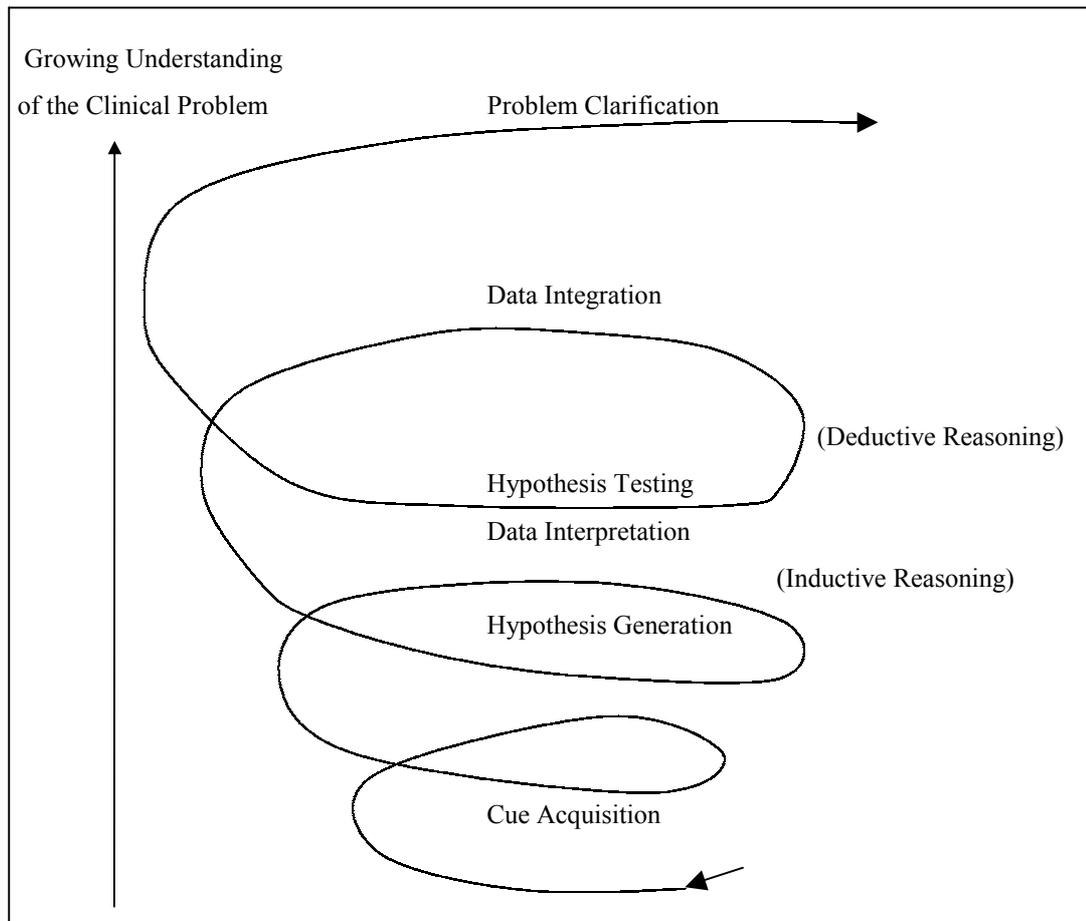


Figure 4.3: Analysis of the Hypothetico-Deductive Reasoning Model
Source: Higgs (1992), Higgs and Jones (1995)

The H-D model of reasoning, while interesting and useful for outlining the CR process, was unsuccessful in delineating differences between expert and novice clinicians (Bordage & Lemieux, 1991). Both novices and experts appear to use this model of reasoning with the difference in performance residing in the organisation of knowledge (Bordage & Lemieux, 1991; Claessen & Boshuizen, 1985; Grant & Marsden, 1987; Newble et al., 1995; Norman et al., 1985; Papa et al., 1990a). This difference in organisation is illustrated in a study by Joseph and Patel (1990). They examined the H-D model in action by studying the verbal protocols of nine doctors (4 endocrinologists and 5 cardiologists) working through a case history on Hashimoto's thyroiditis. The discourse analysis found that the endocrinologists (high domain knowledge) focused on the critical and relevant information. Earlier and more

accurate diagnostic considerations were also generated by this group. While both groups achieved the same diagnosis in the end, the cardiologists (low domain knowledge) generated many more hypotheses and responded to many more cues.

In conclusion, the H-D model of reasoning did not delineate differences in novice and expert performance. The only difference that manifested itself clearly was that experts generated higher quality hypotheses early in the encounter, and this appeared to be a critical factor in determining whether or not they were successful in selecting the correct diagnosis. Norman et al. (1985) suggested that this highly accurate and early hypothesis generation may be the result of a pattern recognition process. This was the view held by Groen and Patel (1988) who viewed the H-D model as being too rational. They felt that experts tended to use forward reasoning and not hypothesis driven inquiry. Barrows and Feltovich (1987), however, continued to support the H-D model, stating that the research did not describe how the patient's pattern is obtained by the doctor in the first place. They felt that gathering data using the H-D model was still needed in order to compare data in the pattern to the actual patient presentation.

Given these controversies in the scientific literature on CR, investigations that attempted to describe the differences in knowledge structure among novices and experts became the focus of subsequent research. The representation of problems in memory, the role of pattern recognition and the influence of biomedical and clinical knowledge became the particular areas of investigation. Before examining these studies in more depth, however, some attention is paid to describing the term, 'knowledge' from the perspective of clinical reasoning.

4.2.2 The Representation of Knowledge in the Memory of Clinicians

What types of knowledge are important in CR, and how is this knowledge represented in memory? Corsini (1984) describes five types of knowledge or learned capabilities from a psychological perspective. A distinction is also made between verbal and motor learning. The five categories are: verbal or declarative knowledge; intellectual or procedural knowledge; cognitive strategies; attitudes; and motor skills. Table 4.2 provides a more in-depth description of these categories.

Table 4.2: Learned Capabilities or Outcomes

-
- Verbal or Declarative Knowledge: this kind of knowledge ranges from single names and facts to bodies of organised information. The kind of performance made possible by such knowledge is stating or declaring information; either orally or in writing.
 - Intellectual or Procedural Knowledge: this kind of knowledge enables the individual, through manipulation of symbols, to demonstrate the application of concepts and rules to specific instances.
 - Cognitive strategies: this kind of knowledge leads to skills which are used to direct and influence cognitive processes such as attending, perceiving, encoding, retrieving and thinking. When these skills are taught to students so they can be deliberately employed, they constitute a major aspect of what is called metacognition.
 - Attitudes: this kind of knowledge comes from affective and cognitive memory components. Attitudes, as outcomes, are learned states that influence the choices of personal action the individual makes towards a person, object or event.
 - Motor Skills: knowledge in this category leads to an outcome which consists of actions accomplished by smoothly timed muscular movements.
-

Source: Corsini (1984)

Three categories of knowledge applied to clinical practice are described by Higgs and Titchen, which take into consideration the work of Polyani, 1958 and Kuhn, 1970 who classified knowledge as propositional and non-propositional (Higgs, 1997; Higgs & Titchen, 1995a; Higgs & Titchen, 1995b). Propositional (declarative) knowledge is derived from research and scholarship and is supported by the professional body. It encompasses book knowledge as well as abstract, logical, and formal relationships between constructs and contexts.

Non-propositional knowledge is divided into two categories (professional and personal) (Higgs, 1997; Higgs & Titchen, 1995a; Higgs & Titchen, 1995b). Professional (craft) knowledge incorporates ‘knowing how’ and the ‘tacit’ knowledge of the profession. It encompasses the practical skills within the profession and is also termed procedural knowledge (Jones, Jensen, & Rothstein, 1995). Benner (1984) has termed this intuitive knowledge, as she believes it develops from experience and personal creativity. Professional craft knowledge guides the

everyday activity of caring for patients and is responsible for the rapid approaches seen in clinician's with expertise (Higgs, 1997; Higgs & Titchen, 1995a; Higgs & Titchen, 1995b). This knowledge, therefore, is tacit and may remain hidden if active reflection does not occur.

Personal knowledge is tied to an individual's reality framework and experience (Higgs, 1997; Higgs & Titchen, 1995a; Higgs & Titchen, 1995b). It is influenced by the personal experiences and reflections of a practitioner and helps them to understand the perspective of their patients. Personal knowledge, such as individual beliefs, values and convictions, also influence propositional and professional craft knowledge (Higgs, 1997; Higgs & Titchen, 1995a). These three forms of knowledge constitute an individual's unique knowledge base and therefore, must be developed as part of the CR process.

In terms of representing knowledge in memory, Christensen (1993) provides an excellent summary of work reported by Anderson, 1990 and Rumelhart and Ortony, 1977. Knowledge is believed to be represented in memory as propositions, propositional networks and schemata. Propositions can be thought of as ideas in memory and are connected to one another through networks wherever there is a relationship. A schema contains networks of ideas or propositional networks. These can be thought of as concepts which allow people to infer information from what they experience. Incoming information from the perceptual systems, therefore, are related to schema in memory, which are in turn activated and retrieved. If new information is received, it is expanded upon using existing schemata with the establishment of more retrieval paths. How this knowledge grows and become organised in novices and experts, therefore, is the next question.

4.2.3 Studies of Novice and Expert Problem Solving in Chess and Physics

Much of the research in CR that followed the H-D model was influenced by studies in chess where research had demonstrated distinct differences in the organisation of knowledge among experts and novices (Chase & Simon, 1973a; Chase & Simon, 1973b; Chi, Feltovich, & Glaser, 1981; de Groot, 1965). In 1965, de Groot studied the chess moves of experts and novices and found that players did not differ in the

depth of processing or in the number of alternative routes explored while engaged in the game. This finding is not dissimilar to the studies of novice and expert clinicians using the H-D model of reasoning. Experts, however, were able to recall chess positions with greater accuracy after observing a chess board for a few moments. A later experiment by Chase and Simon (1973b) supported these results, but also found that if a meaningless organisation of chess pieces was presented to both experts and novices, recall was fairly equivalent across groups. They concluded that chess masters (experts) have better memory and knowledge networks which are confined to their domain of expertise. Thus, it appeared that the organisation of knowledge was a key factor differentiating expert and novice practice.

Larkin et al. (1980) studied differences in experts and novices in solving physics problems and noted that this memory phenomenon in chess was the result of ‘chunking’ familiar information. They defined a “chunk as a cluster of related information units that have become a familiar pattern as a result of repeated exposure”. (p. 1336) These clusters of information are seen as a single unit and expand the memory capacity of an individual. Since short term memory can only retain about five to nine items or chunks of information at any one point in time (Miller, 1956), specific strategies are needed for managing information. In the case of the chess experiments, the chunking process manifested itself in the following manner. The novice can only store one or two chess pieces in a chunk because of their lack of experience. Hence, only 10 to 12 pieces can be reproduced in total. For a master, a configuration of five to six chess pieces may represent a chunk, which explains why they can reproduce the entire configuration of the chess board with much more accuracy. These chunks, as a result, serve as an indexing system enabling the expert to recognise patterns more easily and with more depth.

With respect to physics problems, Larkin et al. (1980) found that experts were able to solve physics problems in one-quarter of the time it took novices, with much fewer errors. The experts also had many automated sequences which appeared to be indexed as patterns. They explained this phenomena by stating:

“although a sizable body of knowledge is a pre-requisite to expert skill, that knowledge must be indexed by large numbers of patterns that, on

recognition, guide the expert in a fraction of a second to relevant parts of the knowledge store. The knowledge forms complex schemata that can guide a problem's interpretation and solution and that constitutes a large part of what we call physical intuition." (p. 1336).

The results of these early experiments in psychology paved the way for investigations in medicine. Carnevali (1995), for example, provides a summary of this 'chunking' process from a clinical perspective. When new information is encountered, it is assigned specific meaning and is encoded. The richness of this encoding is based upon the clinician's knowledge base. This data must be recorded quickly, and mentally rehearsed in working memory, so it is not forgotten. It then ends up being stored in long term memory. This chunking of clinical information, reserves space in working memory and allows clinicians to store more and more pieces of information in each chunk as experience is gained. Experts have more information stored in each chunk and these chunks are used as a strategy to promote pattern recognition.

4.2.4 Pattern Recognition in Medicine

As noted earlier, these early studies of experts and novices in chess and physics became templates for research on CR in medicine. These studies on recall in expert and novice clinicians (Muzzin et al., 1982; Muzzin et al., 1983; Norman, Jacoby, Feightner, & Campbell, 1979) are summarised by Christensen (1993). These studies tried to measure the ability of clinicians with varying degrees of expertise to recall features of written patient case presentations. Cases were typical/atypical representations of disease. In the study by Norman, Jacoby, Feightner, and Campbell (1979) experts could recall more features of the typical case, but recalled similar amounts to novices when the case was atypical - thus replicating the findings of the studies in chess and physics. Muzzin et al. (1982) reproduced this study but standardised the time of exposure to the written case. Experts, in this case, recalled larger chunks or patterns of information indicating a qualitative difference in the organisation of their memory. Muzzin et al. (1983) repeated this study but decreased the time of exposure to the written case. Again, experts were more efficient in their recall as the discrete elements of the case were grouped into categories or patterns. Christensen (1993) notes that these studies reproduced the findings of the

experiments in chess and physics but still failed to differentiate the overall amount of information recalled by novices in comparison to experts.

Grant and Marsden (1987) were one of the first investigators to provide evidence of a consistent difference in the memory structures of novice and expert clinicians. They used the case histories of four real patients and presented them as 10 items of clinical information in a specially designed paper exercise. First (n=15) and third (n=15) year medical students and general practitioners (n=15) were included in the study. The numbers of interpretations (what may be wrong with the patient) and the number of forceful features (items of information that gave rise to the interpretation) did not demonstrate statistically significant differences between groups. However, the actual interpretations (content) in three of the four cases and the actual forceful features in all cases did demonstrate statistically significant differences between groups. There was also considerable variability within a group. Grant and Marsden (1987) concluded that while both groups are able to think broadly to the same degree, the content and organisation of knowledge in the expert group is organised differently.

Other studies that supported the notion of pattern recognition as a means of expert reasoning were criticised for their use of highly visual material, i.e., radiographs or dermatological slides (Groen & Patel, 1988; Norman, Coblenz, Brooks, & Babcock, 1992; Patel & Coughlin, 1985). These researchers argued that this recognition feature may not necessarily apply to complex verbal information that is obtained from patients. Hence, pattern recognition may not apply in all circumstances. In order to explore this possibility further, Groen and Patel (1988) discuss the relationship between comprehension and reasoning in medical expertise. They cite the work of Frederiksen 1975, 1979, 1981 and Kintsch 1974, who used propositional analysis of verbal discourse as a means of studying the perceptual notion of a pattern. By breaking down verbal discourse into units representing recall and inference propositions, a verbal report which is analogous to a pattern can be developed. Using this experimental approach, Coughlin and Patel (1987) and Groen and Patel (1988) report that in routine cases, experts recall more critical cues and make inferences from more highly relevant information. Novices, in contrast, recall and infer more information of low relevance. If information from these routine cases is scrambled,

differences between experts and novices disappear, thus supporting the concept of pattern recognition.

This notion of pattern recognition using verbal discourse protocols is explored in a review of three studies (Groen & Patel, 1988). Some of these studies were re-evaluations of the data originally reported by Muzzin et al. (1982, 1983), however, they were re-interpreted using propositional analysis (Christensen, 1993). Using this approach, it was found that experts who made accurate diagnoses tended to use forward reasoning, whereas those who were inaccurate used forward and backward reasoning. Forward reasoning was used more commonly by experts within their domain of expertise. The superior recall of experts for relevant information, therefore, was explained by the notion that experts know what is relevant because they have accessed a partial model (pattern) from their memory. With atypical or scrambled cases, however, these studies revealed that experts recall as much irrelevant information as novices because the pattern is disrupted.

Coughlin and Patel (1987) also illustrate the influence of case specificity on pattern recognition. They noted that in a routine case with a temporal framework, (bacterial endocarditis), experts recalled twice as much critical information and inferred only half as much information in comparison to novices when the case was presented in a standard clinical format. The familiarity of the case facilitated recall. When the presentation of this case or its temporal framework was scrambled, differences between the novice and the expert diminished, indicating a disruption of the clinician's pattern or prototype. In a less common case (temporal arteriitis), with key findings that point directly to the diagnosis, the experts were less influenced by the presentation order of the case. The experts' accuracy was equivalent in the normal and scrambled case formats. Novices, in comparison to experts, always need to make more inferences because of their lack of knowledge and experience.

This newer approach to CR, pattern recognition, represented a shift from the H-D model of reasoning. It offered an explanation for differences in the reasoning skills of experts in comparison to novices. The terms forward/inductive and backward/deductive reasoning became the terms used to describe the CR skills of experts and novices. These terms are defined in more detail in the next section.

4.2.5 Forward and Backward Reasoning

Forward and backward reasoning is described in great detail in the literature (Arocha, Patel, & Patel, 1993; Patel & Groen, 1986a; Patel & Groen, 1986b; Patel & Kaufman, 1995). Forward reasoning is a data driven strategy in which hypotheses are generated from data using what are termed, 'if...then production rules'. For example, a physiotherapist during the history taking phase of the patient evaluation may find out that the patient has shoulder pain, is a painter and has difficulty elevating their arm. These three pieces of data elicit the pattern, 'rotator cuff tendinitis' in the following way - (IF: shoulder pain, repetitive overhead activity and difficulty elevating the arm...THEN: rotator cuff tendinitis). Hence, when problem cues elicit recognition of the solution/problem, without any specific hypothesis testing, forward or inductive reasoning has occurred.

Forward reasoning requires a sound knowledge base in a particular domain, which is why it is more commonly seen in expert reasoning within that particular area of expertise (Arocha et al., 1993; Patel & Groen, 1986a; Patel & Groen, 1986b; Patel & Kaufman, 1995). While forward reasoning appears to be more efficient, it is also more error prone because conclusions can be made using imprecise or incomplete data (Carnevali, 1995; Higgs & Jones, 1995). These errors in reasoning are described in more detail later in this chapter. The forward reasoning process is represented diagrammatically in Figure 4.4 (Higgs & Jones, 1995). One can see in this model how rapidly reasoning occurs because the intermediate steps of hypothesis identification, testing and data integration, are absent.

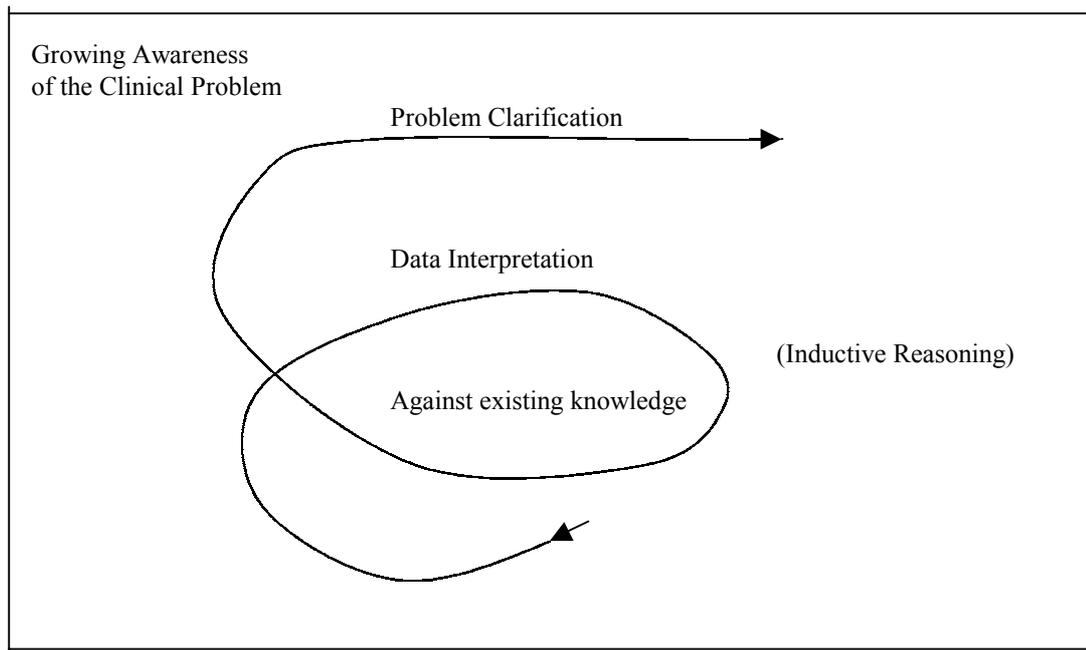


Figure 4.4: Analysis of the Pattern Recognition Model (Forward/Inductive Reasoning)
Source: Higgs & Jones (1995)

Backward or deductive reasoning, in contrast, is a hypothesis driven strategy in which one reasons backwards from a hypothesis and attempts to find data that elucidates it (Patel & Groen, 1986a; Patel & Groen, 1986b; Patel & Kaufman, 1995). It is essentially the H-D reasoning process and is much slower than forward reasoning as it makes heavy demands on working memory because one has to keep track of all the hypotheses. Backward reasoning is more likely to be used when domain knowledge is inadequate and is more commonly used by novices, experts outside of their domain of expertise, or when complex and difficult cases are encountered. Returning to our original physiotherapy example, the same three pieces of information (shoulder pain, painting, difficulty elevating the arm) may elicit several hypotheses in the novice - i.e., rotator cuff tendinitis, sub-acromial bursitis, acromioclavicular sprain and cervical strain. Without the experience to recognise a pattern, the novice must test out each hypothesis by collecting specific data. Data management in this case is much more cumbersome and requires more active intellectual processing.

4.2.6 Organisation of Knowledge as Prototypes

Fueled by the inadequacies of the H-D model to explain differences in the CR of experts and novices, and evidence that knowledge is organised differently in these two groups, Bordage and Zacks (1984) describe a series of experiments that suggest the expert's knowledge is organised around a series of prototypes. These prototypes appear to relate to the concept of pattern recognition. Bordage and Zacks (1984) and Hayes and Adams (1995) argue that traditional views of categorization indicate that all members of a category share common properties, possess the necessary criteria for inclusion in that category and have distinct boundaries that separate them from other categories. This has been found to be incorrect in clinical diagnosis and they note that clinical features that are observed are related instead, in a probabilistic way to categorical labels. For example, they state that no single sign or symptom, in most cases, leads to a diagnosis. These signs and symptoms correlate either strongly or weakly. Through repeated exposure to these category members, therefore, clinician's abstract the central tendency of these exemplars and produce a prototype that contains the features that are most frequently associated with category members. A summary of these features is then stored in long term memory. Future diagnosis, then, involves comparing the features of the presenting case to those stored in a diagnostic prototype.

Bordage and Zacks (1984) state that variability of a feature from the central tendency is quite important as it influences diagnostic prediction. Repeated exposure to cases that vary in their typicality, therefore, fine tune the prototype and promote expertise. This explains why experts are better at perceiving subtle differences or similarities between cases that have similar or different clinical features respectively. Further, it helps to explain why novices are strongly influenced by idiosyncratic features of a case.

As noted earlier, evidence for prototype theory arose from a series of experiments conducted by Bordage and Zacks (1984). They found that disorders with high typicality ratings had the most features in common with other members of that category and were recalled earlier and classified more accurately than less typical disorders. This prototype view also applied equally to both students and general

practitioners. However, more senior clinicians had a richer and tighter network of knowledge. Bordage and Zacks (1984) concluded that prototypes, like patterns, because of their representativeness and overlapping attributes serve as an indexing scheme for the clinician's knowledge. Retrieval of prototypical disorders facilitates recall of other members of the category through the multiple networks within and between categories.

To illustrate this concept in students, Bordage and Lemieux (1986) evaluated the cognitive characteristics of ten medical students with and without diagnostic reasoning difficulties in a problem based curriculum. Using concurrent think aloud procedures as the students worked through a paper case, and reviewing the verbal protocols later, they found that stronger students possessed tighter networks of knowledge which were held together by abstract relationships (semantic axes) such as strong/weak, acute/chronic, superficial/deep etc. This is in keeping with their views of categorization and prototypes. Stronger students explored their knowledge networks and were able to build stronger relationships among the cues. Weaker students without clinical experience used a rote memory strategy comparing signs and symptoms to disorders as they became available. Diagnoses would be added or dropped on a cue by cue basis.

This approach was repeated in a study by Bordage and Lemieux (1991) who investigated 29 second year medical students and 10 specialists using a gastroenterology and neurology paper case. Similar methods of concurrent think aloud analysis, followed by structural semantic analysis of the verbal protocols was utilised. Their findings again reinforced their views of categorization in the form of prototypes. Successful diagnosticians, students or specialists, had more diversified sets of abstract representations and were able to construct broader and deeper representations of the problem. Hence, Bordage and Lemieux (1991) demonstrated that there are even differences within novices and not just between novices and experts. Further, these abstract representations or semantic axes are important components of 'if...then production rules' as they delineate the structure of the rules themselves. For example, IF fever... THEN infection, is one example of a production rule. However, 'fever' can be moderately versus highly elevated, recent versus long-

standing, spiking versus constant and diurnal versus nocturnal. It is these semantic axes that are constructed when learning and organising new information that differentiate successful and non-successful diagnosticians.

Despite the evidence for pattern recognition as a form of reasoning, Barrows and Feltovich (1987) saw pure pattern recognition as too simplistic an explanation for expert reasoning. They argued that clinical problems are generally unknown from the outset so cues must still be obtained by the clinician and these cues are unfolded temporally. Experts with highly organised bodies of knowledge, therefore, are still using highly relevant and tightly competing hypotheses which require only a few inquiries. Data gathering using the H-D model, in combination with pattern recognition, must still take place in order to recognise the pattern, or 'illness script' (Barrows & Feltovich, 1987). Pesut and Herman (1992) address this controversy by noting that the evidence in support of H-D reasoning versus pattern recognition may lie in the methods used to study these phenomenon. The H-D approach often uses live simulations which require Doctors to reason in a way that closely resembles actual practice. With respect to pattern recognition, the case information is often given to the Doctors, which is a lower fidelity situation in comparison to the SP approach. This method of presentation may facilitate pattern recognition as a result.

4.2.7 Illness Scripts

The theoretical notion of an illness script to describe differences in expert versus novice performance was put forward by Feltovich and Barrows (1984) and is described in rich detail by Boshuizen and Schmidt (1995). Illness scripts are composed of three components: enabling conditions; faults; and consequences of the fault. Enabling conditions are conditions or constraints under which a disease occurs and are represented by the personal, social, medical, hereditary and environmental factors that affect health in a positive or negative way, or which affect the course of a specific disease (Boshuizen & Schmidt, 1995). These contextual features are processed into the clinician's understanding of the disease, and when experienced in the clinical environment, activate relevant illness scripts. Faults are the pathophysiological processes that take place in a specific disease and are represented in the script in an encapsulated form. Consequences of the fault are the signs and symptoms

that emanate from the patho-physiological processes. Illness scripts, therefore, are developed from continuing exposure to patients and are a format of rules that enable a clinician to construct a mental model of a family of diseases, a specific disease or even a concrete patient (Schmidt & Boshuizen, 1993). In many ways these are analogous to a pattern or prototype.

Boshuizen and Schmidt (1995) describe the application of illness scripts in experts and contrasts this to novice reasoning. Experts, as part of their CR, search for an appropriate illness script and instantiate it by filling it in with information from the present case. Routine CR, therefore, involves searching for a script, selecting a script and verifying its contents against the present case. Illness scripts are triggered as a whole and provide a list of phenomena to look for in history taking and physical examination. Novices, in contrast, can only rely on their knowledge base and networks, which are much less extensive. They require more information, as a result, before a hypothesis can be created because they only have a few enabling conditions or consequences stored in their memory. Novices, therefore, must reason (backward) through a case in a step by step fashion because their knowledge structures do not automatically produce a script with expected signs and symptoms.

Schmidt, Norman, and Boshuizen (1990) add that illness scripts are highly idiosyncratic because they are constructed, in part, by the doctor's experiential history. Hence, each doctor's script will be different for the same disease, which provides some explanation for the concept of case specificity. Further, these scripts contain relatively little knowledge about the patho-physiological causes of symptoms. Instead, they have a wealth of clinically relevant information about disease, consequences and the context under which the illness develops.

Evidence for the existence of illness scripts can be extrapolated from a study by Hobus, Schmidt, Boshuizen, and Patel (1987). They studied 18 expert and 17 novice clinicians using 32 short case histories which were each presented as a series of three slides. Slide one depicted a portrait of a patient with a neutral facial expression. Slide two was chart information containing previous medical history, psychosocial history and relevant and irrelevant clinical information. Slide three was the presenting complaint and consisted of one to two sentences. Each slide was presented

for a specific duration and after the three slides were shown, the subject had 15 seconds to state the most likely diagnosis. After 16 episodes, the presenting complaint was read back to the candidate and they had to recall the information in the case that gave rise to their hypothesis. These recalls were audiotaped and transcribed for analysis.

They report that experts produced significantly more accurate hypotheses. The total amount of information units that were recalled was also significantly higher in the expert group. A very high correlation was also seen between the total number of accurate diagnoses and recall of contextual information in the experts. This correlation was virtually absent in novices. The researchers explain these findings by stating that experts have better lists of diagnoses stored in their long term memory which are activated by contextual factors. These contextual factors are analogous to the enabling conditions in illness scripts (Feltovich & Barrows, 1984) and illustrate that these contextual features are important in the mental construction of a patient's problem.

4.2.8 An Emerging Theory of Expertise in Medicine

Considering the developments in the CR literature, Boshuizen & Schmidt (1992, 1995), Schmidt and Boshuizen (1993) and Schmidt et al., (1990) present a contemporary theory on the development of expertise in medicine. This four stage theory postulates that expert performance is not due to superior reasoning skills or superior knowledge of patho-physiology, but rather, it is based on cognitive representations in memory that describe the features of prototypical or even actual patients. This theory, which covers the progression from novice to expert, appears to include elements of the H-D model of reasoning and pattern recognition. The theory is described as follows. In stage one, the declarative knowledge of students, acquired in medical school, is transformed into rich causal networks explaining the causes and consequences of disease in terms of general underlying patho-physiological processes. These are called propositional networks and represent how things, objects, events or concepts relate to one another. As learning progresses, these causal networks become increasingly complex.

In stage two, these knowledge networks are compiled and become abridged networks. These causal models, which delineate signs and symptoms, begin to become subsumed under diagnostic labels with repeated exposure to clients. Knowledge that is pertinent to the specific case becomes activated and CR begins to become more efficient. Schmidt and Boshuizen (1993) have studied the reasoning of students at this level. The verbal protocols these students produce are very detailed and contain many biomedical propositions as they are actively processing their abundant basic science knowledge in conjunction with their emerging clinical knowledge. These propositions, however, are often inadequately linked to post hoc explanations of the diagnosis. This increase in reasoning error for clinicians at stage two is called the ‘intermediate effect’ and is representative of the restructuring taking place in the clinician’s biomedical and clinical knowledge.

The progressive shift from a reliance on biomedical knowledge to more clinical forms of knowledge in stage two explains why intermediate students, when shown a case, recall many more details than experts. The lack of this information in experts does not suggest that biomedical knowledge decays with expertise, it is just not used in routine cases. Nonetheless, it still remains available when difficult cases are encountered. For example, experts out of their domain and clinicians who do mostly research, produce more patho-physiological explanations than experts within their domain (Schmidt & Boshuizen, 1993).

In stage three, the emergence of illness scripts occurs. After seeing several similar patients and seeing how disease presentations vary, the patho-physiological networks of stage one and two are compiled into diagnostic and mental labels that explain the phenomena being observed. Emerging experts start to pay attention to contextual factors (enabling conditions) under which the disease occurs. Regehr and Norman (1996) note that these contextual or surface features of the problem are often not perceived by students because of their lack of knowledge and experience. With progressive clinical experience, however, students retrieve and recode this theoretical information into a more useful format so they can see the similarities across concepts or cases. Regehr and Norman (1996) state that students can be assisted here by

having experts explicate the surface features they are attending to when solving the case.

In stage four, patient encounters are stored as instance scripts. Memories of previous patients are retained in memory as individual entities rather than being integrated into a prototype. Experts, therefore, are seen to compare new patients to old patients in memory. These re-collections are synergistic with the illness scripts that are generalisations of a disease. More experienced clinicians, therefore, will likely use their idiosyncratic manifestations of a disease to work through a problem.

The summary of this research suggests that pre-clinical students primarily construct problem representations using detailed biomedical knowledge. This information must be activated in a conscious manner. As soon as patients are encountered, a shift occurs and short cuts are produced. More inclusive concepts emerge which are sufficient for understanding the case and encapsulation and illness script development occurs.

4.2.9 The Influence of Biomedical and Clinical Knowledge

The four stage theory of clinical reasoning expertise outlined above suggests that biomedical knowledge becomes transformed into clinical knowledge and that the two co-exist during the reasoning process. Patel, Groen, and Scott, (1988), however, see biomedical and clinical knowledge as relatively separate entities that are used differently in the CR process. They examined the role of basic science knowledge in CR and had first, second and final year medical students read three basic science texts and one clinical text on bacterial endo-carditis. After reading each text, the students recalled the information, provided a diagnosis and explained the underlying patho-physiology. Analysis of the verbal record was used to examine the recalls. Final year students were better at recalling clinical text and also used more causal explanations to explain their diagnosis. Their causal explanations contained clinical information (knowledge of diseases and associated findings), which was a function of having work experience, and basic science information (anatomy, physiology). There were no trends for basic science recall among the groups but second year students used basic science information extensively, although incorrectly and

inconsistently. First year students used more inference and experiential information given their lack of basic science knowledge. Given that the basic science knowledge was used differently among the three groups, they concluded that basic science and clinical knowledge are used separately and in different ways.

Differences in the use of basic science and clinical knowledge among experts, within and outside their domain of expertise, has also been demonstrated in other studies (Patel, Evans, & Groen, 1989; Patel & Groen, 1986b; Patel, Groen, & Arocha, 1990). These studies suggest that increasing clinical experience is inversely related to the number of biomedical concepts used in explanations of reasoning. These results further support these investigators' claims that basic science and clinical knowledge are used separately.

Boshuizen and Schmidt (1992), however, present an argument that states comprehension of clinical cases comes from biomedical knowledge, which is applied in a more tailored way to the case at hand with increasing expertise. To address this controversy, they conducted two experiments to examine the role of biomedical and clinical knowledge in expert reasoning. In the first experiment the following people were studied: a second year and fourth year student with no clinical experience; a fifth year medical student with some clinical experience (2 rotations); and a family practitioner with four years of experience. Forty eight cards, with a separate clue on each, representing a pancreatitis case were presented serially to each candidate. Their think aloud protocols were analysed to determine whether biomedical or clinical propositions were being presented. Lower level subjects produced more biomedical propositions which decreased with experience. This supported Patel's argument that biomedical and clinical knowledge were used separately. However, qualitative analysis of the texts revealed that the biomedical knowledge was changing as a function of experience.

To increase generalisability, the experiment was repeated with more participants (Boshuizen & Schmidt, 1992). Twenty individuals were studied and spread across the same groups as the first experiment. Subjects were also required to write down the patho-physiological principles underlying the case in this study. Experts were able to show in more detail the biomedical concepts associated with the case and this

explanation was more congruent with their verbal protocol in comparison to novices. Hence, Boshuizen and Schmidt (1992) concluded that biomedical knowledge does not become inert or inaccessible but becomes encapsulated within clinical knowledge. Citing the work on robust knowledge bases by Collins et al. (1989) and Collins (1990), Boshuizen and Schmidt (1992) state that the experts' knowledge base is a product of the active integration of general biomedical knowledge and situated knowledge.

4.2.10 Errors in Clinical Reasoning

As noted earlier, forward reasoning, which is fast and efficient, can lead to errors in reasoning as the data used to come to a diagnostic conclusion can be incomplete. Reasoning using the H-D model can also produce errors. This section provides a brief summary of common CR errors.

Errors in reasoning are often due to errors of cognition resulting from poorly organised knowledge or faulty information processing (Jones et al., 1995; Norton & Strube, 1998). Elstein (1995), Jones (1992), and Watts (1985) provide a comprehensive review of errors commonly made during the CR process in medicine and physiotherapy respectively. These are summarised in Table 4.3. Arocha et al. (1993) note that the psychological research supports the notion that most problem solvers (both lay and scientific) prefer confirmation strategies over disconfirmation strategies. They studied the use of confirmatory and disconfirmatory strategies of second, third and fourth year medical students and one senior resident (n=13). Although the study was small and used only two cases, they found that a confirmatory strategy was used predominately by the groups. Elstein (1995) notes that this is the most common form of reasoning error and occurs as a result of the limitations of working memory. Confirmation bias has also been reported in other health care disciplines outside of medicine (Eli, 1996).

This section of the chapter has sketched out the trends in the CR research, largely from the medical perspective. The advent of clinical experience is seen to be a cornerstone in the pathway to competence as it is at this point that biomedical knowledge begins to become transformed into clinical knowledge. What

distinguishes expert and novice practitioners is not the amount of knowledge per se, although this is still an important factor for successful reasoning, but the manner in which it is organised. The challenge that emerges from the research on CR, therefore, is how to facilitate the development of expertise in the novice. The next sections of this chapter explore the CR literature from the perspective of the allied health professions and the educational implications of CR research in general.

Table 4.3: Clinical Reasoning Errors

Source Elstein (1995)	Source: Jones (1992)
<p>Premature closure: accepting early problem formulation; working through differential diagnoses is a better option</p> <p>Over emphasising rare conditions:</p> <ul style="list-style-type: none"> • sampling bias: over-representing rare conditions • probability distortion: small probabilities are often distorted, particular if they are vague or not precisely known • regret minimisation: the fear of missing a diagnosis that may be harmful may direct clinicians to consider this diagnosis <p>Acquiring redundant evidence: a tendency to seek information that confirms a hypothesis vs. data collection that tests out various hypotheses</p> <p>Incorrect interpretation: most common error is over-interpretation of data that should not support a hypothesis, but it is interpreted as being consistent with the hypothesis under consideration</p> <p>Base rate neglect: balancing out the need to not overlook unusual conditions that are masked as common problems and missing rare conditions that are treatable and potentially harmful vs overcalling rare diagnoses; Avoidance of this error requires specific training in Baye's theorem</p>	<p>Adding pragmatic inferences: making assumptions</p> <p>Considering too few hypotheses: limiting the numbers of hypotheses, thus preventing the correct hypothesis from being uncovered.</p> <p>Failure to sample enough information: making generalisations on limited data, or collecting data in a biased manner based on previous experience</p> <p>Confirmation bias: attending only to those features that support the clinician's preferred hypothesis</p> <p>Errors in detecting covariance: judging the relationship between two factors requires how the two factors covary with one another. Using one combination of covariance to make a decision about the nature of a client's problem may lead to an inappropriate conclusion</p> <p>Confusing covariance with causality: when two factors have been found to covary, it is an error to deduce that these factors are causally related</p> <p>Confusion between deductive and inductive logic: deductive errors would occur in a situation where if A; then B; A is present; therefore B. Inductively this may appear correct but deductively it may not be appropriate as other factors may be causing A which are not B</p> <p>Premise conversion: reversing a statement of categorisation; for example, if A always leads to B, then B must always lead to A</p>

4.3 Clinical Reasoning in the Allied Health Professions

The research on CR in the allied health professions has predominately employed interpretive research designs. The reason for this research design appears to arise from the perspective that reasoning occurs within: the frame of reference of the clinician; the unique knowledge base of the clinician; and the client's specific context (Higgs & Jones, 1995). These aspects of reasoning are generally given less credence in the medical literature on CR and Jones, Jensen, and Rothstein (1995) argue that these components are perhaps more important for clinicians whose tasks require personal involvement in the treatment of clients.

The following sections will look at CR in allied health. A brief review of CR in occupational therapy and nursing will be provided as it provides an interesting contrast to the literature that has been reviewed thus far. Many of these investigations have used qualitative methods in the study of reasoning. A more comprehensive review of the CR literature in physiotherapy follows.

4.4 Clinical Reasoning in Occupational Therapy

Mattingly (1991b) describes CR in occupational therapy as a largely tacit, highly imagistic and deeply phenomenological mode of thinking. Much of this perspective comes from a study on CR jointly funded by the American Occupational Therapy Association and the American Occupational Therapy Foundation (Gillette & Mattingly, 1987). In working with a client, Mattingly (1991a) argues that the occupational therapist sees the client as having the right to choose the direction of their treatment. This is a noticeable departure from the analytical orientation of the CR research in medicine which has guided much of the research on reasoning (Chapparo & Ranka, 1995). Mattingly (1991b) criticises the medical approach for its focus on diagnosis and assumptions that reasoning can be studied in a definite, empirical way. Schon (1991) calls this emphasis on using systematic scientific methods and knowledge to study problems as 'technical rationality', and argues that it limits one's understanding of the artistry of practice. Since the fundamental task of

an occupational therapist is to treat the patient's unique illness experience, Mattingly (1991b) feels that reasoning must be described from a phenomenological perspective.

One of the ways in which the occupational therapist gains an understanding of the patient's illness experience is through narrative reasoning (Mattingly, 1991a). This understanding of the patient's illness experience is critical as the same disease can produce quite different illness experiences in clients. Narrative reasoning, therefore, involves understanding the patient's way of dealing with disability and with puzzling about how to manage the patient (Mattingly, 1991a). This reasoning perspective is comparable to story telling. The story is created through the interaction with the client and leads to an imagined future which then directs treatment. Evidence for the use of story telling is seen in the results of a case study by McKay and Ryan (1995) who studied the reasoning of an occupational therapist with five years of experience and a second year occupational therapy student. Each practitioner told a story about a case. The expert was faster in relating their story, had a future story for the client and was more concise than the novice. The novice adhered to the occupational therapy process, focused more on the biomedical aspects of the case, was disengaged from the process, and had a fragmented story.

An alternative, yet similar view of reasoning is presented by Fleming (1991a, 1991b). Three types of reasoning are described: conditional, interactive and procedural. Conditional reasoning is similar to narrative reasoning and is a composite of interactive and procedural reasoning. Interactive reasoning deals with how the disability or disease affects the client (illness experience) and how the therapist forms a therapeutic relationship with the client (Fleming, 1991b; Neistadt, 1996). It is more commonly seen in occupational therapists with more experience (Schell & Cervero, 1993). Procedural reasoning is closely aligned to the H-D model and is used when thinking about disease and when deciding on treatment interventions. It involves identifying occupational therapy problems and implementing treatment strategies through the use of systematic data collection and interpretation (Fleming, 1991b; Neistadt, 1996). Procedural reasoning is used by all occupational therapists, regardless of the amount of experience they possess (Schell & Cervero, 1993).

In contrast to these more phenomenological approaches, Rogers and Holm (1991) describe reasoning in occupational therapy using the literature from medicine. They apply the H-D model of reasoning to the reasoning process of occupational therapists and assume the generalisability of this CR model. This model of reasoning, which is similar to procedural reasoning, is used when therapists need to consider the patient's problem from a disease perspective within the context of occupational performance (Chapparo & Ranka, 1995). Hence, cue acquisition, hypothesis generation, cue interpretation and hypothesis evaluation are used when occupational therapists need to consider the effects of disease on occupational status (Neistadt & Smith, 1997). The diagnostic reasoning approach described by Rogers and Holm, however, has been criticised by Schell and Cervero (1993) for being based on data from other professions, and for ignoring the personal and practice contexts of occupational therapy. These contexts are important as they activate particular types of knowledge which are specific to a profession (Schell & Cervero, 1993).

Diagnostic or procedural reasoning, however, may still not be diagnostically focused in the traditional medical sense. Fleming (1991a) feels that occupational therapists do not focus on diagnosis, but rather, how the diagnosis will influence the present and future function of their clients. The occupational therapist's focus, therefore, is more individualised to the specific needs of the client which moves them away from the statistical, probabilistic perspective of medical reasoning. Hence, while occupational therapists use a process similar to physicians (H-D reasoning, pattern recognition) when thinking about the client's physical disability (medical problem), they use other forms of reasoning when considering other aspects of the client's life (Fleming, 1991a).

Both Chapparo and Ranka (1995) and Schell and Cervero (1993) provide contemporary reviews of the CR literature in occupational therapy. They describe several factors which influence reasoning in their discipline. These are organisational factors, client factors, the declarative knowledge of the therapist, and the personal values and attitudes of the therapist. Schell and Cervero (1993) also note that there are two strands of reasoning in the occupational therapy literature: scientific (diagnostic) reasoning and narrative reasoning. The former uses logical processes

similar to the H-D model whereas the latter is phenomenological and gives meaning to therapeutic events. They also present a third strand, called pragmatic reasoning, which considers the context in which CR occurs. For example, the treatment environment, the therapist's values, knowledge, abilities and experiences and the therapist's relationship to the treatment possibilities within a given treatment setting (Crabtree & Lyons, 1997; Neistadt, 1996; Schell & Cervero, 1993).

More recent works from Great Britain shed some new perspectives on reasoning in occupational therapy and the analytical versus phenomenological debate (Hagedorn, 1996; Roberts, 1996). Hagedorn (1996), for example, describes the results of a qualitative study that illustrates a strong link between reasoning in occupational therapy to the cognitive psychology literature on reasoning in medicine. She studied the CR of six experienced occupational therapists evaluating a new referral of a common case. The think aloud protocol was recorded and subjected to qualitative analysis. The main form of reasoning used in this study by the subjects was diagnostic/procedural reasoning (Fleming, 1991a; Fleming, 1991b; Rogers & Holm, 1991).

In describing the reasoning of these experts, Hagedorn (1996) discusses the concept of schematic representations. These schema represent stored information which are categorised in memory. Several types of schema are described. The first is a 'dysfunction schema' which is analogous to a prototype or stereotype of the condition. These prototypes are composites of previous knowledge, memories and experiences and are composed as chunks of information. The dysfunction schema is linked to an 'intervention schema', which are memories of all potentially useful forms of interventions, solutions, procedures and tests. These may be stored as 'if...then production rules'. Finally, the 'good outcome schema' contains stereotypes of wellness and recovery as well as expectations of progress and prognosis. Hence, Hagedorn (1996) concludes that in familiar cases, where well constructed schema are available, identification of the problem automatically leads to the identification of a solution which is modified in light of the patient's condition. Hagedorn (1996) terms this predictive reasoning and compares it to the narrative reasoning perspective of Mattingly (1991b).

This description of reasoning, using schematic representations, is comparable in many ways to the literature in medicine which describes reasoning in routine cases. Hagedorn (1996), therefore, concluded that occupational therapists reason like other human beings and other health professionals. This view is certainly in keeping with the those espoused by Newell and Simon, 1972 who see reasoning as following a basic process common to all human beings faced with a problem (Roberts, 1996).

This perspective is supported by Roberts (1996) who is quite critical of the CR research in occupational therapy. Roberts (1996) defines CR from an information processing perspective and sees it purely as problem solving, and not what therapists reason about and what factors influence the way reasoning takes place (phenomenological perspective). Further, she argues that the occupational therapy literature has confused content and process with new forms of reasoning that do not relate to the existing body of knowledge on reasoning. For example, she sees interactive/narrative reasoning as a skill and an important part of what therapists do, but not as a form of reasoning. As far as conditional reasoning is concerned, Roberts (1996) argues that this is the mobilisation of a schema which has been stored in long term memory.

This more recent perspective on the nature of reasoning in occupational therapy, while controversial, is useful as it provides a link between both strands of reasoning. Rather than needing to select one form from another, both perspectives provide useful insights into how occupational therapists work through clinical problems.

4.5 Clinical Reasoning in Nursing

Research on CR in nursing also has employed studies from an analytical and phenomenological perspectives. For example, five levels of expertise in nursing have been adapted from a model originally developed by Dreyfus, 1980 (Benner, 1982; Benner & Tanner, 1987; Benner, Tanner, & Chesla, 1992). This model of expertise was derived from a qualitative study that evaluated the reasoning of 51 experienced nurse clinicians, 11 new graduates and five senior nursing students. These stages are: novice; advanced beginner; competent; proficient; and expert.

The novice reasons through tasks by comparing them to ‘attributes’ and rules learned at school. These attributes can be recognised without clinical experience. However, the use of discretionary judgment is difficult for the novice as the information they have learned usually comes without a context. Hence, information is often applied in a way that may not be appropriate for the real situation. Novices, therefore, spend much of their time testing out their theoretical understandings and seeing how illness states actually manifest in the real situation (Benner et al., 1992).

The advanced beginner has managed enough real situations which enable them to pick up some ‘aspects’. Aspects are global characteristics that require prior experience in actual situations for recognition to occur. Typically, novices and advanced beginners absorb only a little of the experience as they are busy trying to remember attributes and aspects.

Competent, proficient and expert skill is illustrated by more holistic practice, with the nurse being able to see the attributes and aspects of the current and future situation. Proficient practice is guided by ‘maxims’, which enable the nurse to see a situation completely. These maxims teach nurses what to expect in a situation and how to modify plans quickly if the expected picture varies. Experts, in contrast, possess an intuitive grasp of the situation which is quick and accurate (Benner & Tanner, 1987). The use of past concrete experiences guide practice, instead of rules and formulas, and the recognition of patterns assists nurses to predict what lays ahead (Benner et al., 1992).

While these characteristics of CR in nursing are described in a qualitative manner, they parallel the research on CR in medicine and the theory of expertise proposed by Schmidt and Boshuizen (1993). For example, attributes can be likened to the reliance on biomedical knowledge in novice reasoning, whereas aspects are analogous to the causal networks or prototypes used by novices with some clinical experience. Maxims, which are used by proficient nurses, are similar to illness scripts and the use of past concrete experiences to guide practice in expert nurses resemble instance scripts. Hence, while this description by Benner and colleagues is useful from a nursing perspective, it illustrates the potential generality of the reasoning process across the health professions.

Corcoran (1986a, 1986b) also looked at the planning skills of five novice and six expert nurses when managing straightforward and complex cases. Three written cases, of increasing complexity, involving the management of a client with pain in a hospice setting were used. Think aloud procedures were employed along with protocol analysis to evaluate the nurses' reasoning. There were no statistically significant quantitative differences between expert and novice nurses in the number of alternative actions generated. From a qualitative perspective, however, the experts' alternatives were described in more detail and they had more explicit rationales for their evaluations. Further, alternatives (or hypotheses) were often generated before they were evaluated. These findings are similar to other studies of novice-expert reasoning and provide support for the use of H-D models of reasoning in nursing (Barrows & Feltovich, 1987; Elstein, Shulman, & Sprafka, 1978; Fonteyn, 1995; Jones, 1988; Neufeld, Norman, Feightner, & Barrows, 1981).

From a qualitative perspective, Corcoran (1986a) describes three types of knowledge that nurses use in their reasoning. These are: declarative or factual knowledge; procedural knowledge; and experiential or tacit knowledge. In the study described in the preceding paragraph, experts were seen to demonstrate more declarative, procedural and experiential knowledge and generated more alternative and detailed pain management strategies. Further, experts were seen to chunk information as they could retain complex drug protocols by creating analgesic ladders. This increased the efficiency of the nurses' short term memory capacity in comparison to the less experienced nurses who had to remember each drug separately. Again, these findings parallel other studies on expertise in medicine, chess and physics, where chunking of information is seen to increase the efficiency of the information processing system (Chase & Simon, 1973a; Chase & Simon, 1973b; de Groot, 1965; Larkin, McDermott, Simon, & Simon, 1980; Muzzin et al., 1983; Muzzin et al., 1982; Norman, Jacoby, Feightner, & Campbell, 1979).

Itano (1989) has also studied the clinical judgment process of 13 expert and 13 student nurses. She used real patients (n=26) in a medical surgical specialty area and audiotaped the nurse-patient encounter. Subjects reviewed their thinking while listening to the audiotape. This thinking aloud record was also captured and

transcribed, along with the actual interview. Despite the lack of patient control in this study, Itano found that the type and number of cues did not differ significantly between the groups although the experts tended to collect more cues. Hence, while judgment differed between the groups, the process and scope of cue acquisition did not appear to be different. These results provide further support for the generality of H-D reasoning processes.

Fonteyn (1995) provides a comprehensive review of the CR literature in nursing. She notes that a variety of studies using information processing theory, decision tree analysis and phenomenology have been implemented to study the reasoning of nurses. From a nursing perspective, she describes CR as:

“the cognitive processes and strategies that nurses use to understand the significance of patient data, to identify and diagnose actual or potential patient problems and to make clinical decisions to assist in problem resolution, and to enhance the achievement of positive patient outcomes.”
(p.60).

Fonteyn (1995) differentiates this from the nursing process and cites work by Phillips and Rempusheski 1985, and Hurst et al 1991, who found that the linear nursing process of assessment, diagnosis, planning, implementation and evaluation, is not necessarily followed by nurses when working through a clinical problem.

Fonteyn (1995) points out that there is evidence to suggest that nurses use the H-D reasoning model as well as the forward reasoning approach associated with pattern recognition. For example, Benner and Tanner (1987) illustrate that more experienced nurses are less likely to use analytical processes in their reasoning and employ what appears to be intuition to grasp the clinical situation. This intuitive reasoning appears to be equivalent to many other forms of reasoning in the literature which are aligned to other professional disciplines. For example, Benner and Tanner (1987) note that intuitive reasoning is composed of six sub-components which all work in unison. They are: pattern recognition, similarity recognition, commonsense understanding, skilled know-how, sense of salience and deliberative rationality. Pattern recognition occurs when relationships are observed between concepts which subsequently lead to a pattern that can guide treatment. Similarity recognition involves recognising the

characteristics of a current case to one that was encountered in the past. These two features of intuition, when combined, appear to resemble the notion of pattern recognition described in the medical reasoning literature (Groen & Patel, 1988; Muzzin et al., 1983; Muzzin et al., 1982; Norman et al., 1979).

In conclusion, while Benner's notion of intuitive reasoning is interesting, it parallels many forms of reasoning described in the literature. The research on CR in nursing also suggests that there is a strong relationship between analytical and phenomenological paradigms. For example, information processing theory, or the analytical perspective, seeks to explore the contents and actions of the reasoners' mind whereas phenomenological skills acquisition theory seeks to explore reasoners' experiences of reasoning (Greenwood, 1997). These perspectives emanate from differences in experimental method, the philosophical approach of the discipline and the theoretical background of the investigator carrying out the research.

4.6 Clinical Reasoning in Physiotherapy

Unlike medicine, the amount of research on CR in physiotherapy is sparse and most of it is in the area of orthopaedics and manual therapy. Relying on the research in medicine to describe the CR of physiotherapists can be problematic as the practice patterns of the physiotherapy profession are different. For example, physiotherapists evaluate the patient problem, identify factors that are amenable to treatment and attempt to effectively manage the problem (Higgs, 1990; Jones, 1992). This interventionist focus is quite different from the approach taken by medicine. Rothstein and Echternach (1986) support this view by stating that physiotherapists just do not think of a diagnosis but need to understand how the limitations are affecting the patient and the impact it has on function.

4.6.1 Diagnosis and Physiotherapy

While questioning the applicability of diagnostic practice, in the medical sense, to physiotherapy, it would appear that physiotherapists still make diagnoses. Delitto and Snyder-Mackler (1995), for example, review several perspectives on the issue of diagnosis in physiotherapy. They found that physiotherapists do not identify disease in the sense of pathology, but diagnose movement-related dysfunctions. By clustering

data from a variety of sources, a classification system that is similar to a diagnosis is used to direct the management of the problem. Several definitions of diagnosis from a physiotherapy perspective are described as they have relevance to this study since subjects will be required to develop a 'diagnosis'.

Sahrmann (1988) defines diagnosis as:

“the term that names the primary dysfunction toward which the physical therapist directs treatment. The dysfunction is identified by the physiotherapist based on the information from the history, signs, symptoms, examination and tests the therapist performs or requests.” (p. 1705).

Umphred (1995) describes physical therapy diagnosis as:

“the process used by therapists to (1) analyse the patient’s specific impairments, (2) how those component problems interact with larger bodily system functions, and (3) how disabilities are created from those specific problems. The process is intricately intertwined with selecting appropriate evaluation and treatment procedures.” (p. 39).

This definition is similar to one offered by Jette (1989) who sees physiotherapists constructing diagnoses to guide their treatment. He argues that medical diagnosis is very linear with a focus on identifying a singular disease, a perspective that is too narrow for physiotherapy.

These perspectives suggest that the generation of diagnoses do occur in physiotherapy, although they are more multifaceted and include both medical and social aspects. Requiring the entry level practitioner to generate diagnoses, however broad, is arguably an entry level competency.

4.6.2 Hypothetico-Deductive Reasoning in Physiotherapy

Jones et al. (1995) conducted a comprehensive review of the literature on CR in physiotherapy. They feel that CR is influenced by the therapist’s perspective, the patient and the practice environment. Therapists are also seen to employ the H-D model of reasoning in their practice, using their interventions and reassessments as supporting evidence for their hypotheses (Grant, Jones, & Maitland, 1988; Jones et

al., 1995; Umphred, 1995). The CR process in physiotherapy, therefore, is seen to be very similar to the process described in the medical education literature (Grant et al., 1988). However, Higgs (1992) states that physiotherapy places a greater emphasis on treatment and subsequent evaluation rather than diagnosis. Jones (1992, 1995) specifically describes six categories of hypotheses that physiotherapists use as part of their reasoning. They are: sources of symptoms or dysfunction; mechanism of symptoms; contributing factors; precautions and contraindications to physical examination and/or treatment; management and treatment; and prognosis. Hypotheses can also be physical, psychological or socially related (Jones, 1995). Rivett and Higgs (1997) have introduced a seventh hypothesis termed, “reassessment”.

From these descriptions of reasoning, one might conclude that physiotherapists observe and interpret initial cues which lead to working hypotheses; collect data as part of the subjective and physical examination process with questions and procedures tailored to the specific patient and the therapist’s hypotheses; accept or reject their various hypotheses depending upon what they find during the examination; and continue with this process until a diagnostic or management decision is reached (Jones, 1992; Jones, 1995; Rivett & Higgs, 1997).

This model of H-D reasoning is evident in earlier descriptions of the clinical decision making process of physiotherapists (Grant et al., 1988; May & Newman, 1980; Rothstein & Echternach, 1986). While these descriptions are theoretically based, and not the result of formal experiments, they describe several sequential steps that closely parallel the H-D reasoning model. May and Newman (1980) describe a seven step process: problem recognition; problem definition; problem analysis; data management; solution development; solution selection and implementation; and outcome evaluation. Rothstein and Echternach (1986) describe a nine step process: collect initial data; generate a problem statement; examination or collection of data; generate working hypotheses; plan re-evaluation methodology; plan treatment strategy based on hypotheses; plan tactics to implement the strategy; implement the tactics; and reassess results. While both of these models possess elements of the

original H-D reasoning model described by Elstein et al. (1978), they go on to illustrate the role of the physiotherapist beyond diagnosis.

Experimental evidence for the use of the H-D model of reasoning in physiotherapy is also evident (Edwards, Jones, Carr & Jensen, 1998; Payton, 1985; Rivett & Higgs, 1997). In Payton's study, methods similar to those used by Elstein et al. (1978) were employed, except that real patients were used in his study and the retrospective verbal protocols of the physiotherapists were stimulated using audiotape. Ten physiotherapists were studied. The range of experience was from three to 18 years with expertise cutting across several different specialty areas. The results of this study suggest that physiotherapists use a CR process that is similar to the one used by medical practitioners.

Rivett and Higgs (1997) studied the CR of 11 expert and 8 less expert manual therapists. All subjects independently observed a videotape of a clinician performing a history on a patient with a neuromusculoskeletal problem. Subjects concurrently recorded their thoughts on data collection forms at 7 predetermined intervals. The frequency and categories of hypotheses were then measured and compared across the two groups. They found that both groups generated hypotheses early in the encounter and used the H-D process. There were also no statistically significant differences between the two groups in terms of the number of hypotheses generated across categories.

Edwards, Jones, Carr, & Jensen (1998) studied the CR of expert physiotherapists in three areas of physiotherapy (manipulative/orthopaedic, neurological and domiciliary care) and not only found that physiotherapists generated and tested hypotheses, but also revealed the use of pattern recognition processes. This qualitative study also demonstrated that expert physiotherapists employ several of the CR strategies that have been described in the nursing, occupational therapy and physiotherapy literature. For example, diagnostic/procedural reasoning (Fleming, 1991b; Jones, 1988; Payton, 1985), conditional or predictive reasoning (Fleming, 1991b; Hagedorn, 1996), narrative reasoning (Benner et al., 1992; Mattingly, 1991a) and ethical/pragmatic reasoning (Schell & Cervero, 1993).

4.6.3 Phenomenological Approaches to Reasoning in Physiotherapy

The review of literature provided by Jones et al. (1995) does not limit itself to the H-D reasoning model. They also support the phenomenological approach, as they see physiotherapists also having personal involvement in the treatment of their clients. Understanding the patient's perspective is an important part of the physiotherapist's reasoning as this influences the choice of management possibilities. A variety of studies illustrate this perspective (Christensen, 1993; Embrey, Guthrie, White, & Dietz, 1996; Jensen, Shepard, Gwyer, & Hack, 1992; Jensen, Shepard, & Hack, 1990).

Jensen et al. (1990) studied eight physiotherapists in outpatient settings managing clients with musculoskeletal problems. Only one therapeutic session per client was studied and a total of 19 treatment sessions were observed in all. Observation, field note records and audiotape recordings of therapists engaged in practice were the methods used for the study. This data was then qualitatively analysed for emerging concepts. Jensen et al. (1990) cite the work of Roter (1988) and report that the therapists' communication patterns were very similar to medical practitioners. An extra category, instructing and teaching the patient, however, was added to make their analysis more pertinent to physiotherapy. They found that the more experienced clinicians spent more time with their clients on hand-on activities, information seeking and evaluation/education. The more experienced therapists could also manage interruptions more easily. Therapists with less experience tended to focus on data gathering whereas the more experienced clinician could engage themselves in social interchange, information giving and treatment simultaneously. While this study does not describe the CR process per se, it does provide some insight into the work of the physiotherapist and the differences between the expert and novice practitioner.

A second study by Jensen et al. (1992) used similar investigative procedures to study the practice of three master clinicians and three novice clinicians in an outpatient setting. Five attribute dimensions of practice that were distinguishable between master and novice clinicians were identified. These attribute dimensions were: ability to control the environment; evaluation and use of patient illness and disease data;

focused verbal and non-verbal communication with the patient; equal importance of teaching to hands on care; and confidence in predicting patient outcomes based on knowledge of pathology and experience with the course of healing. Jensen et al. (1992) note that the masters had a more elaborate cognitive framework with more easily accessible schemata to guide their evaluation and treatment. Masters also had more procedural knowledge, which the investigators argue is more dynamic than declarative knowledge. Master clinicians, therefore, were able to rule items in and out until a collection of signs and symptoms were collected and corroborated. Novices, in contrast, were less selective of their data collection and gathered as much information as possible.

The results described by Jensen et al. (1990, 1992) present in qualitative terms, some of the features of CR described in the medical education literature. For example, evidence of pattern recognition and forward reasoning can be seen in the practice pattern of the master physiotherapist whereas the backward reasoning approach appears to be more common in the novice physiotherapist. The more elaborate cognitive frameworks of the master physiotherapists also supports the notion of prototypes and illness scripts as methods which guide practice in the expert (Boshuizen & Schmidt, 1995; Feltovich & Barrows, 1984; Hayes & Adams, 1995; Schmidt et al., 1990). Umphred (1995) supports these results by noting that expert performance in physiotherapy is explained in part by the experts' ability to match their memory and experiences of cases to the actual patient case before them. This helps the physiotherapist formulate the direction of their assessment and intervention.

Similar results are described by Embrey et al. (1996) who studied the clinical decision making of three experienced and three inexperienced physical therapists in paediatrics. Retrospective think aloud procedures were employed with the verbal text of the protocols analysed to demonstrate differences in reasoning. The inexperienced therapists possessed fewer, 'movement script schemata' than the more experienced therapists and were less likely to use these schemata automatically or spontaneously. The inexperienced physiotherapists also had less procedural changes during therapy and relied on lists of activities to guide the therapy session. They also found that experts were more spontaneous and paid more attention to psychosocial issues

whereas the inexperienced group were more activity oriented. Attention to these psychosocial or contextual features may be similar to the concept of enabling conditions (Feltovich & Barrows, 1984). These enabling conditions facilitate the retrieval of illness scripts and therefore enhance the reasoning efficiency of the expert. Hobus et al. (1987) noted that contextual features were used by experts to activate their knowledge networks in a streamlined manner, thus reducing the amount of processing needed to work through a case. The higher attention paid to psychosocial issues and patient illness data in the studies by Embrey et al. (1996) and Jensen et al. (1992) respectively, suggest that similar categorisation processes may be used by expert physiotherapists to guide their practice in routine or familiar cases.

4.6.4 Pattern Recognition in Physiotherapy

Grant et al. (1988) note that pattern recognition is used by physiotherapists as part of their reasoning and illustrates this using a case study of a patient with 'shoulder' pain. They feel that the emergence of pattern recognition (or forward reasoning) emanates from the combination of biomedical knowledge and clinical experience. Experienced physiotherapists, by virtue of their more organised knowledge base, are able to recognise syndrome patterns more readily. This enables them to process incoming cues much more discriminately.

This descriptive illustration of pattern recognition by Grant et al. (1988) has been more formally investigated by Christensen (1993), who looked specifically at clinical pattern recognition in a small group of physiotherapists. Five written cases describing different low back pain syndromes were presented to two groups of physiotherapists with different levels of experience in musculoskeletal practice. Subjects were asked to correctly recognise and identify the clinical pattern in terms of the most likely diagnosis and to record the production rules that they used in identifying the given clinical pattern. The production rules that were described by the two groups were compared and contrasted within and across the groups as well as to the production rules established by a group of three expert physiotherapists. Four types of production rules were evaluated which classified information into the following categories: contributing factors/predisposing factors; site and/or nature of symptoms; history; and behaviour of symptoms. Three types of production rules were

also described that characterised whether subjects were reasoning using confirming statements, negating statements and hypothesis limiting statements.

Christensen (1993) found no statistically significant differences between groups in the average number of production rules containing the four different types of information or in the average number of production rules that were representative of the types of reasoning statements. These findings are in keeping with the study by Grant & Marsden (1987) who found that both experts and novices were able to reason through a problem with similar breadth and that both groups employ a H-D model of reasoning.

In terms of differences between the groups, Christensen (1993) found a positive correlation between actual experience and low back pain pattern recognition ability for physiotherapists with at least eight years of experience. No such correlation occurred for physiotherapists with less than two years of experience. This supports the claims in the literature that diagnostic accuracy increases with expertise (Groen & Patel, 1988; Patel & Groen, 1986).

Significant intra-group variation was also seen in the breadth of production rule usage for the group with more experience. Again, this is not surprising as expertise is highly idiosyncratic (Grant & Marsden, 1987; Schmidt et al., 1990). There was also a positive trend for the experts to rely on the use of contributing factor/predisposing factor information. Christensen (1993) notes that this preference is comparable with the literature on illness scripts. These studies show greater use of enabling conditions in expert reasoning (Feltovich & Barrows, 1984; Hobus et al., 1987; Schmidt et al., 1990).

Both the novice and expert groups demonstrated a greater reliance on the use of confirming statements to reason through the cases (Christensen, 1993). Again, this is in keeping with the literature on errors in CR and the universal tendency for individuals to use confirmatory data to support their reasoning (Arocha, Patel, & Patel, 1993; Eli, 1996; Jones, 1992). Experts, however, were also seen to use other types of reasoning statements (i.e., negating statements and hypothesis limiting statements) more frequently than novices. This suggests that experts were using

more forward reasoning processes and were more discriminatory in making their interpretations.

This small study by Christensen (1993) provides support for many of the CR processes in medicine being applicable to physiotherapy and illustrates some of the differences in novice and expert practice. In terms of limitations that reduce the generalisability of this study by Christensen (1993), only a small non-randomised sample of physiotherapists from one educational program was studied. Lack of consistency between markers in scoring the number and quality of expert production rules that were cited by subjects was also evident. This inconsistency obviously influenced the reliability of the data. Further, the distinction between the two groups in terms of their actual experience was not large with some overlap between the two groups. This would limit one's ability to demonstrate clear differences between the groups.

4.6.5 Clinical Reasoning of Physiotherapy Students

Very few studies in the physiotherapy literature actually examine the CR processes of entry level physiotherapy students. Isles (1995) studied the clinical decision making processes of three strong and three weak final year physiotherapy students in neurology. Observation and audiotaping of the students' practice was carried out by the investigator. Students were also stopped at intervals and asked to retrospectively reflect on their thought processes. Students were found to rely on backward reasoning although two of the strong students were seen to use some forward reasoning. Strong students were also more reflective and accurate in their identification of cues. Poor students made more errors of omission, were less accurate and were more routine in their questioning and examination.

Thomas-Edding (1987a) compared the clinical reasoning of 24 expert clinicians (>3 years experience) with 24 novices (students who had completed their academic course work and had completed at least one clinical placement). Four SP cases, two in orthopaedics and two in neurology, were used. Within each specialty area, one case was considered simple and the other complex. Each subject's SP encounter was videotaped and afterwards, the subjects were required to produce a list of prioritised

problems as well as a treatment plan. The videotapes were coded using broad categories to measure the time spent on various aspects of the patient encounter. These six categories for the orthopaedic cases were: introduction and history taking; testing for movement/strength/joint play; sensation; palpation; swelling; explanations/instruction/summary. Transcripts of audiotapes made from the videotapes were also analysed from a qualitative perspective.

Thomas-Edding (1987a) reports largely on the results from the orthopaedic cases (excision of a right scaphoid and neck/shoulder pain of unknown etiology). Experts spent significantly more time with the orthopaedic patients in all six categories in comparison to the students. This is an interesting finding as it is in contrast to the literature that generally depicts experts as being faster (Glaser & Chi, 1988). Thomas-Edding (1987b) also noted that experts spent much of their evaluation time taking a complete and thorough history, carefully attended to highly relevant cues and tried to establish a global picture of the condition. The process she describes is very similar to the narrative/conditional reasoning processes described in the occupational therapy literature (Fleming, 1991a; Fleming, 1991b; Mattingly, 1991a). Novices did not engage in this form of reasoning to the same degree and merely applied a routine interview structure with standard questions. This approach appeared to be used by the students to compensate for their lack of knowledge. Since the students did not respond to cues to the same degree as experts, this may have been the reason for their shorter encounter.

In terms of the H-D model as a form of reasoning, experts had fewer hypotheses than the students. These hypotheses were also more general in comparison to the more specific hypotheses of the students. While both parties were seen to employ forms of H-D reasoning, Thomas-Edding (1987b) argues that experts relied more so on data driven inquiry methods. This approach is similar to the forward reasoning process (Groen & Patel, 1988). Students' hypotheses, in contrast, were more numerous and were generated by lists of signs and symptoms which they used as checklists to work through the patient problem. While these checklists items resulted in a higher percentage of hypotheses, they also resulted in the students ignoring pertinent patient cues.

The results reported by Thomas-Edding (1987a) clearly illustrate differences in physiotherapy student and expert reasoning. Her results also support the claim that expertise is characterised by a well developed and organised knowledge base. Novices were seen to rely on their biomedical knowledge to work through the case, a finding that parallels stage one of expertise as described by Schmidt et al. (1990). Further, Thomas-Edding, (1987b) notes that the expert physiotherapists were not seen to be using a completely intuitive approach to their reasoning, which is different from what the medical education literature reports on expert medical diagnosis. The need for the physiotherapist to construct a detailed and personal history in order to understand the full scope of the patient's problem requires additional time. Hence, this may explain why experts took longer to work through the cases in comparison to the students.

The preceding sections have provided a review of the CR literature in the allied health professions. Hypothetico-deductive and pattern recognition forms of reasoning appear to be used by most health professionals depending on their level of expertise and familiarity with the case at hand. The phenomenological reviews of CR, however, do provide rich descriptions of the reasoning process beyond the cognitive approaches used in the medical education literature and are nonetheless valuable for increasing one's understanding of the CR process.

4.7 Clinical Reasoning and the Novice Practitioner

4.7.1 The Novice: A definition

The term, novice, has been used in the literature to describe clinicians with varying degrees of experience. This study sees the novice as an undergraduate physiotherapy student with a good biomedical knowledge base but little clinical experience. Higgs and Edwards (1999) describe the beginning practitioner as being technically and professionally competent; equipped with generic skills such as self-directed learning skill; and having the ability to interact effectively within their environment. Oldmeadow (1996) also provides a definition from the student perspective. "The novice student is one who lacks clinical experience. Knowledge, psychomotor and affective skills have been learnt previously, but not in the context in which they must

now be applied.” (p. 39). These definitions are relevant for this particular study as they describe practitioners who have adequate knowledge but very little, if any, clinical experience.

4.7.2 The Novice Practitioner: Clinical Reasoning Characteristics

This section provides a summary of novice performance. One thing that is very clear in the literature is that novices rely on their biomedical and applied science knowledge to reason through a clinical problem (Boshuizen & Schmidt, 1992; Boshuizen & Schmidt, 1995; Newble, van der Vleuten, & Norman, 1995; Schmidt & Boshuizen, 1993; Traband & Dunn, 1988). This knowledge base is poorly organised in comparison to experts (Barrows & Feltovich, 1987; Boshuizen & Schmidt, 1995; Jensen, Shepard, Gwyer, & Hack, 1992; Norman et al., 1985; Regehr & Norman, 1996; Traband & Dunn, 1988). For example, novices are seen to use more biomedical propositions when reasoning through a case, despite the fact that these propositions are often poorly linked to their post-hoc explanations of their diagnosis (Schmidt & Boshuizen, 1993).

Jones (1992) in a review of the literature, notes that novices recall more surface features of a problem and have fewer patterns of conditions stored in their memory. Hence, novices must rely on their biomedical knowledge for reasoning, using textbook cases as their patterns. When cases vary from this simplistic prototype, novices experience difficulty in working through the case as they have not integrated their declarative and procedural knowledge into meaningful patterns (Jones, 1992; Regehr & Norman, 1996). For example, Whelan (1988) found that the majority of pre-clinical medical students used what he called, an “ordering approach” to solve problems. This atomistic perspective was characterised by students ignoring symptoms, discussing symptoms individually and not organising groups of relevant information into meaningful wholes. Explanations relied heavily on the use of biomedical knowledge.

When novices begin to see patients, however, their biomedical knowledge and clinical experiences are used to transform their knowledge into more meaningful clusters of information thus facilitating more efficient and accurate reasoning

(Carnevali, 1995; Schmidt & Boshuizen, 1993). This is seen in experiments where clinicians with varying degrees of experience are asked to describe their reasoning when given a case (Groen & Patel, 1988; Patel & Groen, 1986). Novices generally recall more relevant and irrelevant information whereas experts make more inferences from highly relevant cues. This latter feature demonstrates a much deeper understanding of the problem and a more organised knowledge base among experts. Novices, in contrast, because of their poorly organised knowledge base, and lack of clinical patterns, must rely on backward reasoning strategies to a much larger degree (Jones, 1995; Jones, Jensen, & Rothstein, 1995).

Novices also are seen to have lower levels of procedural knowledge in comparison to experts as much of this knowledge is acquired through experience (Jensen et al., 1992; Jones et al., 1995). This lack of procedural knowledge has an impact on how students manage a case. For example, Oldmeadow (1996) states that without having any history of situational contexts in which to apply their knowledge, students find it difficult to judge the relevance and importance of certain aspects of a task. Hence, all features of a clinical assessment will be given equal attention and significance.

Novices also use H-D forms of reasoning, however, their hypotheses tend to be narrow in contrast to experts who keep them more broad (Barrows & Bennett, 1972; Thomas-Edding, 1987b). This is because each cue that is offered by the patient generates specific hypotheses. These hypotheses are based on biomedical principles and not clinical patterns (Barrows & Feltoich, 1987; Boshuizen & Schmidt, 1995; Higgs & Jones, 1995). With respect to the number of hypotheses generated, novices appear to be able to generate the same breadth of responses to clinical information as their more experienced counterparts (Grant & Marsden, 1987; Neufeld, Norman, Feightner, & Barrows, 1981). However, the hypotheses of novices do not contain the richness of information that are seen in experts (Elstein, 1995). Further, these hypotheses are not as detailed when it comes to descriptions of the cause and nature of functional problems and treatment possibilities (Chapparo & Ranka, 1995).

Novices without experience and those with very little experience are also less likely to respond to broader psychosocial cues, are less spontaneous in their actions, are task oriented and tend to be guided by lists or frameworks (Embrey, Guthrie, White,

& Dietz, 1996; Fonteyn, 1995; Higgs & Jones, 1995; Jensen et al., 1992; Jensen, Shepard, & Hack, 1990; Thomas-Edding, 1987a; Thomas-Edding, 1987b; Tichenor et al., 1995; Traband & Dunn, 1988). Novices have also been found to put less emphasis on patient education and teaching in comparison to hands-on activities (Jensen et al., 1992).

In conclusion, the novice practitioner is in many ways the opposite of an expert. Gilhooly (1990) describes several key features of the expert. These maxims Gilhooly outlines are useful for understanding how novices differ from experts. These maxims are: experts remember better; experts work forwards; experts have better problem representations; experts have a superior knowledge base but not necessarily a better capacity for knowledge; and experts become expert through extensive practice. These same maxims have been summarised by Glaser and Chi (1988) and are presented in Table 4.4 below.

Table 4.4: The Nature of Expertise

-
1. Experts excel mainly in their own domain because of their knowledge within that domain. (Expert Doctors have more differentiations of common diseases so they can more readily see variations in comparison to novices).
 2. Experts perceive large meaningful patterns in their domain. (This not due to more superior perceptual abilities but due to the organisation of their knowledge base.)
 3. Experts are faster than novices at performing the skills of their domain and quickly solve problems with little error. (This is due, in part, to extra practice and the stored condition-action rules in which a specific pattern or condition triggers a stereotypical sequence of moves.)
 4. Experts have superior short term and long term memory. (The automaticity of many portions of their skills frees up resources for greater storage.)
 5. Experts see and represent a problem in their domain at a deeper level than novices who represent them more superficially. (Experts categories are semantically or principle based whereas the categories of novices are syntactically based or organised around surface features.)
 6. Experts spend a great deal of time analysing a problem qualitatively, at the start of problem solving whereas novices plunge immediately into the problem. (Experts build a mental representation from which they can infer relations that define the situation and they add constraints to the problem; novices do not see the constraints.)
 7. Experts are more aware than novices when making errors. (Experts have superior monitoring skills which reflects their greater domain knowledge as well as their different representation of knowledge.)
 8. Experts have a depth of understanding of the clinical problem which includes the client's perspective (This particular citation is an addition by Higgs & Jones (1995)).
-

Source: Glaser & Chi (1988)

4.8 Educational Implications of the Clinical Reasoning Literature

The research on CR provides numerous perspectives on how to improve teaching and learning in novice health professionals. Carr, Jones, and Higgs (1995) provide an excellent description of an educational outcome for physiotherapy programs and provide justification for developing teaching and learning strategies that augment reasoning. They state that:

“courses must produce graduates who are capable of sound problem analysis and client management within their area of expertise, and who are able to recognise and refer on to others problems which are outside that area. ... They (graduates) need to be able to set their practice within recognised social, ethical and legal mores, and work cooperatively and communicate effectively with others” (p. 235).

Schon (1983) has noted that experienced clinicians can perform problem solving rapidly. Novices, unfortunately cannot. Hence, learning experiences, such as peer

assisted learning (PAL) methods, can be used to promote the patterning that is needed for more timely problem solving. Methods for achieving these outcomes are the focus of this section, taking into consideration the perspectives of the research on clinical reasoning.

4.8.1 The Patient Evaluation Process

Barrows (1990) and Barrows and Tamblyn (1980) make a strong argument for encouraging students to generate early hypotheses in their patient encounters. Citing the research on the H-D model of reasoning, they, along with others argue that reliance on menu driven inquiry methods may be inappropriate (Benner & Tanner, 1987; Delitto & Snyder-Mackler, 1995; Gale, 1982; Leighton & Sheldon, 1997; Terry & Higgs, 1993). These menu formats encourage students to collect comprehensive data without early hypothesis generation. The end result is that students collect reams of data that is unrelated to the specific condition or hypotheses. This only serves to confuse the students. Encouraging students to develop early hypotheses and then organising their data search around these hypotheses is more appropriate as this is how clinicians solve problems in the real setting.

Menu driven inquiry, using the subjective, objective, assessment, plan (S.O.A.P.) format that is common in physiotherapy has been criticised (Delitto & Snyder-Mackler, 1995). They see educators confusing completeness with exhaustiveness when this inquiry method is being taught. The end result is that educators often teach laundry lists of examination procedures that the students are somehow expected to perform in a proficient and strategic manner. While menu driven search methods are useful for preventing reasoning errors, as they ensure students do not forget to ask key questions, they should take a secondary role (Barrows, 1990; Gale, 1982; Terry & Higgs, 1993). Problem directed inquiry, instead of menu driven searching methods, should be developed using case studies and problem based learning methods. Barrows (1990) argues that this focus encourages the students to apply the cognitive skills they will need to apply in the future setting.

4.8.2 Experiential Learning: A Constructivist Perspective For Teaching and Learning in Physiotherapy

The use of experiential learning methods appear to be a key teaching strategy for the development of CR skill in novices. Henry (1989) describes several activities that encompass experiential learning. Some of these are the use of group discussion, case studies, journals/diaries, work and community placements, simulations, and problem based learning. Discussion of all of these activities is beyond the scope of this thesis. However, some attention will be paid to group discussion as this relates specifically to the concept of reciprocal peer coaching (RPC), the focus of this thesis.

Experiential learning is an educational philosophy, a range of methodologies and a framework for being, seeing, thinking and acting on individual and collective levels (Refshauge & Higgs, 1995). May and Newman (1980) also point out that effective CR is most likely to occur in an environment where the student is free to test out their thinking skills, explore alternatives and to discover alternatives that may or may not match other clinician's solutions. These perspectives certainly support more experiential approaches to learning.

The impact of experiential learning methods on the cognitive structure of the learner is described by several authors (Barker-Schwartz, 1991; Boud, 1988; Boud & Edwards, 1999; Brown, Collins, & Duguid, 1989; Graham, 1996). Boud (1988, 1993) criticises educators for placing little emphasis upon how students can learn from complex experiences outside the classroom. Far too much energy is placed on learning outputs and not inputs. Given that constructivists see learners as constructing their own knowledge it is argued that more attention needs to be paid to learning from experience (Boud, 1993; Brown et al., 1989).

A variety of fundamental experiential learning principles describing how students learn, are described in the literature (Boud, 1993; Boud & Edwards, 1999). They state that experience is the foundation of, and the stimulus for, learning. Further, prior experiences influence all learning as learners bring the totality of their life history to the learning setting. Learners also construct their own experience, being influenced by affective, cognitive and conative features as well as the social and cultural context. Learning is also seen to occur within a socio-emotional context with

input from groups and peers. Brown et al. (1989) describe this type of learning, which encompasses both the physical and social contexts, as situated learning. Learning in these authentic situations allow concepts to evolve because the new situation, and the negotiations/discussions that occur, recast the information into a more densely textured form (Graham, 1996).

Learning partnerships are one strategy that can maximise the experiential learning experience (Boud, 1988; Boud & Edwards, 1999). Students share experiences, clarify expectations of the placement, provide mutual support and ask questions of one another which they might be inhibited from asking their supervisor. Wynn and Kromrey (1999) have noted that when feedback is provided by a directing teacher, the person being observed may be overly concerned that the feedback is evaluative rather than supportive.

Boud (1988) and Boud and Edwards (1999) provide an experiential model for learning partnerships. The first stage involves 'returning to the experience' or 'prior practice' so that the learner can recapture as many parts of the experience as possible. The second stage involves 'attending to feelings' or 'current practice'. Here, the learner examines what feelings were evoked. The learning partner is particularly useful here as they can point to areas that have been misinterpreted or unexamined. 'Re-evaluating the experience' or 'following practice is the final stage. Here, the new experience is related to prior experiences and new knowledge is re-organised using a variety of cognitive and metacognitive strategies: association; integration; validation; validity testing; and appropriation.

The influence of experiential learning is also seen in the theoretical perspectives of educators, particularly those who encourage the use of discussion as part of the learning process. Barker-Schwartz (1991) cites the work of Belenky et al. (1986) who describe two concepts: connected knowing; and separate knowing. Connected knowing is a preferred educational orientation that includes the sharing of common experiences and discussion of the feelings that inform ideas. Separate knowing is an orientation to learning that is characterised by impersonal and objective reasoning, commonly referred to as critical thinking. Barker-Schwartz (1991) argues that

learning activities that involve discussion of experiences, that illustrate theory in practice, will promote connected knowing.

This perspective on learning is also supported by others (Boshuizen & Schmidt, 1992; Carr et al., 1995). They see the integration of general and situated knowledge emanating from reflection and discussion. By exploring the connections between these two knowledge domains, encapsulation of biomedical knowledge into more relevant and robust clinical forms can take place. Further, by transforming biomedical knowledge into more useful clinical formats, transfer to future encounters may be enhanced.

Specific strategies for promoting the transformation of knowledge into clinically relevant forms are described in detail (Carr et al., 1995; Cohn, 1989; Graham, 1996; Higgs, 1990; Higgs, 1992; Neistadt, 1996; Refshauge & Higgs, 1995; Terry & Higgs, 1993). The need to encourage this transformation emanates from the medical education literature. For example, Boshuizen and Schmidt (1992) state that encapsulation of biomedical knowledge is the second stage of a four stage model denoting expert practice. Strategies that encourage this knowledge transformation, therefore, are needed. Tutorial groups that explore a variety of cases and use the H-D reasoning approach to work through problems, are recommended as one teaching strategy for pre-clinical students. Learning from mistakes and exploring alternative treatment decisions can be the focus of discussions in these groups, which may enhance transfer of learning to the clinical setting. This learning format also encourages discussion among peers, clarifies reasoning processes and introduces metacognition into the learning experience.

For students engaged in clinical practice, discussion about patient problems with a peer can also be used to encourage knowledge transformation. This can be facilitated by exposing students to the same patient over the course of a placement and having them see other patients with similar/dissimilar diagnoses (Anderson, Reder, & Simon, 1996; Cohn, 1989; Grant, Jones, & Maitland, 1988). The discussion that emanates from these experiences should enable students to create stronger relational structures in their knowledge base, leading to better encapsulation of their knowledge.

The development of a robust or encapsulated knowledge base is well supported in the literature when experiential learning methods are applied. This perspective is supported by Gale (1982), who cites the work of Crutchfield (1972). Crutchfield states that:

“if we seek to nurture the student’s ability to think, then we must give him appropriate training on the many specific skills we have described. But in order to do this most effectively... we should train them simultaneously in the context of whole problems which have considerable scope, complexity and meaningfulness. In this fashion, the student will practise using his productive mental processes in the integrated way they must be used for genuine problem solving.” (p. 196).

The use of experiential strategies such as peer discussion and collaboration, therefore, can foster this link between biomedical knowledge and the context of real clinical practice. The utility of peer discussion in developing reasoning skill, therefore, is explored in more detail in the next section.

4.8.3 Peer Discussion

Numerous researchers support the concept of discussion with peers as one strategy for the promotion of CR skills (Barker-Schwartz, 1991; Boshuizen & Schmidt, 1992; Boshuizen & Schmidt, 1995; Graham, 1996; Higgs, 1992; Refshauge & Higgs, 1995; Regehr & Norman, 1996; Schell & Cervero, 1993; Schmidt, 1983; Terry & Higgs, 1993). Peer discussion can be incorporated into tutorial based settings with multiple students as well as in the clinical setting where only two students, for example, may be working together. The communication that takes place between peers exposes the learners’ thoughts and arguments and allows for discussion and restructuring of knowledge to take place. This forces learners to explicate their reasoning and fosters the development of metacognitive skill since it requires students to think about how they think, and to consider how much they know or do not know.

One of the reasons peer discussion may elevate comprehension is seen in a study by Bordage and Lemieux (1986). They studied the CR skills of students in a problem-based curriculum using a text-based case and found that the stronger students had more highly developed knowledge structures with stronger relational semantic

networks. Weaker students relied on syntactical knowledge structures, which resulted in fewer and weaker mental operations during the CR process. Encouraging students to discuss cases amongst themselves, therefore, may actually enhance their reasoning because of the opportunity to review, revise, and add to their existing relational knowledge networks. This view is supported by Resnick (1988) who sees collective problem solving as a convenient way of accumulating knowledge. Resnick also sees collective problem solving leading to insights and solutions that would otherwise not occur as it brings to light misconceptions that have been erroneously directing practice.

Peer coaching may also enhance the information processing capabilities of the learning team. Barrows and Tamblyn (1980) note that cue perception is limited by the capacity of the human senses. Only so much information can be perceived at any one point in time by a single person. Hence, translation errors can occur during reasoning whereby a symptom or complaint is missed or misinterpreted. The presence of a peer during the patient encounter, however, provides another set of perceptual tools and also provides the learner or novice with a human sounding board when they are unsure about the relevance of patient information. Novices can also be encouraged to think aloud during patient encounters to reveal their reasoning to the other novice(s) (Corcoran-Perry & Narayan, 1995). The thinking that is revealed can be used to identify where reasoning errors have occurred, provided the other novice has the ability to pick these up.

Peer coaching may also be a more efficient way of providing feedback and supervision to novices in the clinical setting. Clinical instructors may not be available to provide frequent feedback to the students because of their patient care responsibilities. This lack of feedback may increase learner stress (Oldmeadow, 1996). Peer coaching, therefore, is one means of providing additional support and feedback to the students. Students, collectively, can work through simple problems and resolve them for themselves. In those situations where they cannot answer the question completely, they can approach their supervisor for assistance on those aspects of the problem they find difficult.

The benefit of peer discussion, therefore, appears to lead to an enhanced awareness that arises from discussing biomedical and clinical information. This awareness can be likened to metacognition, a cognitive skill important for the management of knowledge. The influence of metacognition is examined in more detail in the next section, with reference to peer discussion when appropriate.

4.9 Metacognition and the Management of Knowledge

4.9.1 Definitions of Metacognition

In the previous section it was noted that the discussion that takes place between peers enhances metacognition as it requires students to think about how they think, and to consider how much they know or do not know. The importance of metacognition from the perspective of reasoning, therefore, is described more fully in this section.

Metacognition has been defined in a variety of ways, making its interpretation broad and vague (Pesut & Herman, 1992; Strohm-Kitchener, 1983; Worrell, 1990). For example, terms such as cognitive monitoring, self-communication, metamemory, metacomprehension, learning strategies, and study skills have been used to describe metacognition. Higgs and Titchen (1995) and Jones (1995) define metacognition as being aware of one's own cognitive processes and controlling them. Further, they feel that good metacognitive skills lead to deeper learning, as reflection-in-action and reflection-about-action (Schon, 1991) is what leads to expertise. Reflection-in-action is thinking about what you are doing while you are actually doing it. Reflection-about-action is quite similar, but occurs after the task or experience is complete. In both situations, the learner or clinician engages in a critical analysis of the problem which leads to adaptations or alterations to their practice. Pesut and Herman (1992) define metacognition as the self-communication one engages in, or the internal dialogue that one emits before, during and after performing a task. Hence, they see metacognition including such things as knowing what one knows, knowing when and how one comes to know it, being able to think and plan strategically, the ability to represent knowledge effectively and in ways that permit efficient retrieval, and the ability to monitor and consistently evaluate one's own competence.

Metacognition is also seen to be a critical skill by Nickerson, Perkins, and Smith (1985) who feels that expertise is influenced by the ability to manage one's intellectual resources..." (p.68). This theme of 'managing' your knowledge, is expanded upon by Jones et al. (1995) who state there is a strong relationship between biomedical knowledge, cognition and metacognition. Jones states that, "without time to reflect ... many clinicians will fail to learn new clinical patterns and their reasoning and practice will increasingly be based on clinical routines with little relationship to biomedical knowledge" (p. 23).

Flavell (1979) was one of the first people to formally describe the concept of metacognition. He sees it playing an important role in memory and problem-solving, and describes it in terms of 'metacognitive knowledge' and 'metacognitive experience'. Metacognitive knowledge consists of knowledge and beliefs about how certain factors (person, task and strategy) interact in ways to affect cognitive outcomes. Each factor has inherent variations. For example, people have their own beliefs and tasks come with sufficient/insufficient information and are familiar/unfamiliar. These variations tell the problem-solver how successful they will be in achieving their goals. Hence, metacognitive knowledge, leads you to select, evaluate or abandon cognitive tasks, goals and strategies. Metacognitive experiences, in contrast, are any cognitive or affective experiences that pertain to an intellectual enterprise. They can be brief/lengthy, simple/complex and are likely to occur in novel situations where a lot of planning and evaluation occur. Novel situations offer numerous opportunities for thoughts and feelings to emerge about one's own thinking. Hence, metacognitive experiences lead to new goals or to the abandonment of old goals and add to one's knowledge base.

Strohm-Kitchener (1983) argues that there is one more step, above metacognition, which she terms epistemic cognition. She notes that most experiments that have described metacognition have used puzzles, which are well structured problems with known solutions. The definitions that have come about from these studies have led to metacognition being understood in a very broad way, that is, it represents the monitoring of all cognitive phenomena. However, most adults do not spend their time solving puzzles, but rather, are faced with ill-defined problems which are often

rife with conflicting assumptions, evidence and opinion, all of which can lead to different solutions. Clinical problems, particularly for the novice, would fall into this category. Working through ill-defined problems, according to Strohm-Kitchener, therefore, requires the use of epistemic cognition.

A three tiered model of cognitive processes in ill-defined problems is described by Strohm-Kitchener (1983), which includes this epistemic cognitive perspective. Each tier provides the framework for the next tier. The first tier represents straightforward cognitive tasks such as reading, memorising and perceiving. The second tier represents metacognition which encompasses the processes that are used to manage the cognitive processes of the first tier. Metacognitive processes include knowledge about cognitive tasks, i.e., how to memorise, strategies that can be used to solve the task, when or how the strategy should be applied, and the potential success or failure of the strategy. These first two tiers do not appear to differ substantially from conventional definitions of metacognition. The third tier represents epistemic cognition which is the process the individual uses to monitor the nature of the problem and the truth value of alternative solutions. It includes the individual's knowledge about the limits of knowing, the certainty of knowing and the criteria for knowing. Hence, when working through an ill-structured problem, the person must decide if the problem is solvable, how it is solvable and whether there are strategies to solve the problem.

4.9.2 The Influence of Metacognition on Learning

Joyce and Weil (1996) support the importance of metacognition in learning from a constructivist perspective. Constructivists believe that learners actively construct their own reality or knowledge using personal experience to structure their rules, concepts, hypotheses and associations (Biehler & Snowman, 1997). This personalised knowledge is influenced by the learner's experiences, beliefs, gender, age, ethnic background and individual biases (Biehler & Snowman, 1997; Graham, 1996). Joyce and Weil (1996) challenge the view that knowledge and skill development must be provided to students in neat packages. Instead, they see knowledge development as personal, and occurring within a social perspective through interactions with teachers and other learners. Hence, each learner's concepts

and meanings will be unique, and the learner's task, therefore, is to seek meaning within their own unique frame of reference. Building knowledge, and checking it against the concepts of others, is seen by to be a major part of the process of education (Biehler & Snowman, 1997; Joyce & Weil, 1996).

The importance of metacognition in learning is also seen in an overview of adult learning theory (Mezirow, 1981). He applies the theories of Jurgen Habermas (1971, 1979) to his critique and illustrates how personal awareness about knowledge enhances cognition. Mezirow discusses three cognitive perspectives in learning which influence the generation of knowledge: technical, practical and emancipatory. Mezirow argues that most educational methods emphasise the first two perspectives, which focus on the provision and evaluation of knowledge and skills. While these methods may be appropriate for competency based evaluation, Mezirow feels they ignore the emancipatory perspective. Emancipatory learning:

“involves an interest in self-knowledge, that is, the knowledge of self-reflection, including interest in the way one's history and biography has expressed itself in the way one sees oneself, one's roles and social expectations. ... Insights gained through critical self-awareness are emancipatory in the sense that at least one can recognise the correct reasons for his or her problems” (p. 5).

This type of metacognitive knowing is critical, particularly if one considers the importance of pragmatic or ethical reasoning (Jones, 1997; Neistadt, 1996; Schell & Cervero, 1993). These forms of reasoning involve considering the moral, political and economic dilemmas in clinical practice.

Emancipatory learning can be promoted by encouraging discussion and dialogue with peers and by participating in and leading learning groups (Mezirow, 1981). This helps the learner to identify real problems involving, for example, power relationships, institutional ideologies that are embedded in myths, and their own personal feelings. Mezirow feels that by critiquing these psycho-cultural perspectives, alternative meaning perspectives can be created.

More contemporary views of metacognition, and their influence on learning have also been described in the literature (Biggs, 1987). He refers to ‘meta-learning’ as a specialised application of metacognition. Meta-learning develops in the early teens when students begin to become aware of their motives for learning and begin to select, deploy and control specific learning strategies in order to realise their motivational intentions. These motives determine what strategies (surface versus deep) will be used as part of the learning. Surface and deep approaches to learning are described by Biggs (1987, 1988) and are summarised in Table 4.5.

Table 4.5: Approaches to Learning

-
- Surface Learning: Where the motive to learn is to meet institutional requirements minimally with the congruent strategy of using a reproductive strategy through rote learning to meet the essentials.
 - Deep Learning: Where the motive to learn is intrinsic. There is interest in the content and the strategy is to discover meaning by reading widely and inter-relating this information with existing knowledge.
 - Achieving: Where the motive to learn is ego enhancement through high grades and the strategy is organising time, space and coverage of content.
 - Surface Achieving: Where the motive is to achieve but accurate reproduction of detail is seen as the approach for success
 - Deep Achieving: Where the motive to learn comes from an intrinsic interest and a desire for high grades. Approaches to work come from an organised and strategic search for meaning.
-

Source: Biggs (1987, 1988)

Ramsden (1985, 1988) provides an excellent overview of the research on student learning, tracing the development of deep and surface approaches to learning from the early work of Marton and Saljo (1976) to the more recent theoretical perspectives of Biggs (1982, 1983) and Entwistle (1983, 1985). The definition of the deep approach to learning has been broadened over time and is seen as an orientation to learning that is internally focused. The student with a deep learning approach is concerned with the content and structure of the task and on integrating the meaning of the task with previous knowledge and experience. Learners with a surface approach, in contrast, tend to focus on task requirements. They rely on memorisation and fail to distinguish principles from mere facts. Hence, their learning is externally focused.

From an educational perspective, Ramsden (1985, 1988) states that surface and deep approaches to learning are influenced by assessment methods and the amount of content that is provided to students. For example, excessive content and assessment practices that encourage rote memorisation lead to surface learning approaches. In contrast, student-centered approaches that reduce anxiety and demonstrate concern for the learner facilitate deeper learning. Hence, mechanisms that encourage students to confront the discrepancies in their present way of thinking are more desirable. Discussion with other learners who are learning the same subject matter is one option that can be used to promote deeper learning.

Certain tasks, particularly those that are complex, may also benefit from the application of metacognitive teaching strategies. Learning tasks that involve high structure, high fact ratios (S/F ratio) are described by Biggs (1987). He states that tasks with high S/F ratios require more complex performance as a lot of information has to be reproduced and comprehended. This would certainly be the case in a novel patient encounter where the initial problem is poorly defined and requires a depth of performance to find a solution. Biggs (1987) states that tasks with a high S/F ratio require deeper approaches to learning. If this is the case, the use of metacognitive strategies to enhance clinical performance would be useful. Biggs (1988) suggests the use of peer teaching, self-questioning and group based techniques as metacognitive strategies that can be factored into the teaching context.

This section on the novice practitioner and the educational implications of research on CR has attempted to highlight the potential usefulness of peer coaching/discussion strategies. The importance of metacognition as a factor in promoting deeper learning was also described in considerable detail as it is seen as a central skill for the development of expertise. Peer discussion is one potential way of enhancing the amount of metacognitive deliberation during learning and is of major interest to the investigator.

4.10 Qualitative Research Methods to Study Clinical Reasoning

The remaining section of this literature review focuses on methods to study clinical reasoning.

4.10.1 Using Verbal Reports to Study Clinical Reasoning

Documenting the cognitive and metacognitive strategies of learners is a growing research activity in many fields and there are a variety of methods that can be used to study this reasoning (Ericsson & Simon, 1993; Garner, 1988, Yinger, 1986). These methods essentially involve having the subject verbalise their thinking about a task. The verbal record or protocol that is captured, becomes the piece of data used for analysis. One of the more common verbal data methods that have been used to study CR in the health sciences is stimulated recall (Mast, Feltovitch, Soler, & Myers, 1985). Stimulated recall enables an investigator to study what a clinician does in a relatively naturalistic patient encounter by capturing the clinician's reasons for his/her actions and interpretations. It is also useful in situations where concurrent verbalisation of one's thinking activities are difficult to obtain (Yinger, 1986); for example, when treating a patient.

4.10.2 Theoretical Foundations for the Use of Verbal Reports as Data

The theoretical framework which supports the use of verbal data is based upon information processing theory (Ericsson & Simon, 1984; Ericsson & Simon, 1993). This theory asserts that humans process information using two distinct memory systems: short and long term memory (LTM). The short term memory (STM) system can only process five to nine items or chunks of information at any one point in time (Miller, 1956). The contents of STM, unlike the LTM, are immediately accessible in the minds of individuals as these information chunks are being processed at the conscious level.

Ericsson and Simon (1984, 1993) state that by studying STM, investigators can learn about the cognitive processes subjects use to solve problems. Inferences can also be made about what is stored in the subject's LTM (Ericsson & Simon, 1984; Ericsson & Simon, 1993; Henry, Le Breck, & Holzemer, 1989; Jones, 1989). To illustrate how this is possible, Ericsson and Simon (1984, 1993) argue that the cognitive processes which occur during problem solving generate numerous verbalisations which are in themselves a subset of the cognitive processes that generate any type of recordable response or behaviour. Those information processes that are currently being concentrated upon, as well as those that were recently acquired, are directly

accessible because they are being processed in the short term memory. Hence, by having subjects verbalise their thoughts in response to a problem solving task, their verbal record reveals the contents of their short term memory. The completed verbal report, becomes a record of the sequential steps used to solve the problem.

Verbal reports can be captured in several ways and have been categorised into taxonomies (Ericsson & Simon, 1984). The categories are based upon the time of verbalisation in relation to the cognitive task, and the relationship between heeded and verbalised information. Level one verbalisations are elicited by having subjects think aloud or talk aloud while attending to a task (Ericsson & Simon, 1993). Thinking aloud occurs without prompting whereas talking aloud occurs with prompting. Level one verbalisations represent the subject's internal speech and are assumed to be direct verbalisations of what is stored in STM (Patel & Arocha, 1995). Both thinking or talking aloud verbalisations are obtained when the subject is verbalising concurrently with the cognitive task. The relationship between heeded and verbalised information is considered to be high (Ericsson & Simon, 1984; Yinger, 1986).

Because of the prompting that occurs in talking aloud, the verbalisations in this category contain some information from the long term memory (Patel & Arocha, 1995). In particular, information that is linked to what is being processed in the short term memory. Ericsson and Simon (1984) state that there is still a high relationship between heeded and verbalised information when talking aloud is taking place.

Level two verbalisations are descriptions of recently acquired information which have been influenced by cognitive processes such as scanning, filtering and inference (Henry et al., 1989). Ericsson and Simon (1993) see level two verbalisations arising from a process called concurrent probing. Here, the subjects are prompted by the investigator to keep on talking and to respond to queries while they are engaged in a cognitive task. The relationship between heeded and verbalised information is still high but the verbal records are a mixture of information in the STM and LTM (Ericsson & Simon, 1993; Yinger, 1986).

Level three verbalisations emanate from retrospective verbal reports (Ericsson & Simon, 1993; Yinger, 1986). Here, the subject is asked to report what they were thinking about after engaging in a cognitive task. There can still be a relatively high relationship between heeded and verbalised information if the recall is done immediately after the task as this minimises demands on the subject's long term memory (Garner, 1988). Garner (1988) and Nisbett and Wilson (1977) note that separation in time between the verbal report and the actual occurrence of the process is the greatest factor decreasing accuracy. Garner (1988) provides evidence for this by reporting on a study in which subjects verbalised their strategies after engaging in a summarisation task. Subjects either recalled their activities on the same day or two days later. It was found that subjects who recalled on the same day reported significantly more cognitive events than the delayed report subjects. Yinger (1986), however, argues that information still disappears from the STM within minutes of the original event. Hence, level three verbal data is likely to be a combination of recent experience (STM) and of information stored in LTM such as similar memories. The relationship between heeded and verbalised information, therefore, is not as high as in level one and two verbalisations.

The difference between level three verbalisations and those verbalisations captured in stimulated recall, however, are not explicitly discussed by Ericsson and Simon as they assume subjects are not re-supplied with cues when verbalising their thought processes (Yinger, 1986). In light of this difference, Yinger (1986) provides an excellent description of how stimulated recall relates to the principles espoused by Ericsson and Simon (1984). According to Ericsson and Simon, only information in the STM, or information which provides associative cues to episodic memory in LTM can be directly reported. Researchers who use stimulated recall assume that the cues on videotape are isomorphic to the original cues and, therefore, allow subjects to retrieve information residing in LTM (Yinger, 1986). Hence, the verbalisations that occur during stimulated recall influence thinking in the same way as during the original event.

Yinger (1986) argues, however, that the videotape review is, in itself, a new event even though it is related closely to the original episode. In light of this, the subject

and the researcher are unable to differentiate whether the verbalisation is the result of direct recall or whether it has been reconstructed. Nevertheless, because the verbalisations are closely related to the original event, Yinger (1986) compares the verbalisations that occur in stimulated recall to those that occur during concurrent probing (level 2 verbalisations). The fundamental difference, though, is that researchers are probing the subjects' STM for the new videotape event and accessing links to the LTM that are related to the original event.

4.10.3 Criticisms and Limitations of Verbal Report Data

Several researchers have noted that verbal data may be inaccurate, particularly if the methods used to capture this data are not sound (Garner, 1988; Nisbett & Wilson, 1977; Smith & Miller, 1978; Yinger, 1986). Nisbett and Wilson (1977) particularly criticise the use of verbal reports and state that researchers are unable to gain access to the direct thoughts of any individual. This stems from their doubts that people have the ability to observe directly the working of their own minds. Hence, any attempt to access information on higher order cognitive processes will lead to information that is less than accurate.

Smith and Miller (1978), however, disagree with the arguments put forward by Nisbett and Wilson (1977). For example, they note that many of the experiments that Nisbett and Wilson (1977) use to argue against the use of verbal report data are considered to be 'uninvolving' to subjects. Highly automated or overlearned skills may not be reported by subjects because of their 'uninvolving' nature (Garner, 1988; Smith & Miller, 1978). This automaticity would under-represent the integrity of the verbal report and would explain why subjects do not report on the intervening cognitive processes.

With respect to retrospective verbal reports, distortion can occur because of new thinking that arises following the encounter (Elstein, Shulman, & Sprafka, 1978; Gilhooly, 1990; Higgs, 1992; Yinger, 1986). For example, subjects are often unaccustomed to seeing themselves on videotape and may focus on what they are doing rather than what they were thinking at the time (Yinger, 1986). They may also notice things that they did not perceive during the original event which may influence

their verbalisations. In light of this, the videotape event may be perceived quite differently from the original event. Verbal responses may also be subject to various outside influences such as social desirability, evaluation apprehension and demand characteristics (Smith & Miller, 1978), although one could argue that this is possible in both concurrent and retrospective recall situations.

Methods to ensure that retrospective verbal data are accurate, however, have been described by Elstein et al. (1978). They note that distortion can be detected by carefully observing the commentary of the subject. If the subject makes a comment about a cue that has not yet occurred, or which incorporates information that was received after the event, this is distortion and the evidence is removed.

4.10.4 The Influence of Verbalisation on Task Performance

Jones (1989) has noted that thinking aloud during practice is unnatural and not part of a clinician's normal repertoire. Hence, comprehensive data collection could be compromised by the numerous tasks that have to be done during patient care. Jensen, Shepard, Gwyer, and Hack (1992) also found that novice physiotherapists were less able to control the treatment session when other environmental influences got in the way. Further, in a study by Greenwood and King (1995), one nurse stated that some of her attentional capacity while treating a patient was focused on the requirement to think aloud.

Fonteyn and Fisher (1995) and Henry et al. (1989), however, suggest that concurrent verbalisations do not influence cognitive processes or effect the speed of task performance as long as the subject is not queued by the investigator. One problem associated with this approach, however, is the fact that if one minimises the amount of cueing to maximise the integrity of the cognitive process, incomplete records of the metacognitive and cognitive activity may result (Garner, 1988). This occurs because the lack of prompting often results in fewer verbalisations. It would seem that the choice of verbal data collection that one chooses is dependent upon the objectives of the study. For straightforward tasks and/or tasks of short duration, concurrent verbalisations may be appropriate. Tasks that are more complex or require a fair bit of time to complete may be more appropriate for retrospective

recall. The richness of the data that one would like to obtain, and the concomitant amount of probing that would be required to elicit this data, may also dictate whether concurrent or retrospective verbalisations are obtained.

4.10.5 Stimulated Recall and Studies of Clinical Reasoning

Elstein et al. (1978) were one of the first groups of researchers to use stimulated recall to investigate clinical reasoning. They used introspective reports to study the CR of 24 doctors in order to determine the thought processes, heuristics and decision rules used to solve a problem. They note that the use of introspective research has a long history in psychological research (Bloom, 1953; Kagan & Krathwohl, 1967; Peterson & Clark, 1978; Siegel et al., 1963) with this method dating back as early as 1934 (Fonteyn & Fisher, 1995).

Stimulated recall is also a valuable tool for analysing the nature of the CR process of students (Barrows, 1987; Mast et al., 1985). This allows the evaluator to assess reasoning, hypothesis generation, inquiry strategies, data analysis, problem synthesis and diagnostic/therapeutic decision making. Stimulated recall can also be used as a non-directive technique in conjunction with the use of simulated patients as a method for studying clinical reasoning (Barrows & Feltovich, 1987). This approach, they argue, is one of the best and most sophisticated attempts at modeling clinical reasoning.

Both concurrent and retrospective (stimulated recall) verbalisation procedures have been used in studies of clinical reasoning. Table 4.6 provides a more in-depth summary of these studies. As can be seen from this table, both approaches (depending upon the objectives of the study) are widely used as an investigative method. Small sample sizes are generally employed because of the massive amount of data that is generated by each recall episode (Fonteyn & Fisher, 1995; Fonteyn, Kuipers, & Grobe, 1993). For example, for every 30 minutes of videotape, approximately 60 - 90 minutes of recall takes place (Barrows, 1987; Elstein, Kagan, & Shulman, 1972; Elstein et al., 1971; Elstein et al., 1978). Significance testing is also not the objective of these recall sessions, which is why small sample sizes are typically investigated. Instead, investigators are concerned with capturing rich

qualitative descriptions of the CR process. Methods for analysing this qualitative data are described in the next section of this chapter.

Table 4.6: Studies Employing Retrospective (R) or Concurrent (C) Verbalisations for the Study of Clinical Reasoning

Reference	Discipline	R/C	Methodology
(Arocha, Patel, & Patel, 1993)	Medicine	C	Paper case in cardiology
(Barrows, Norman, Neufeld, & Feightner, 1982)	Medicine	R	Simulated patients
(Bordage & Lemieux, 1986)	Medicine	C	Paper case: neurology
(Bordage & Lemieux, 1991)	Medicine	C	Paper cases in gastroenterology and neurology
(Boshuizen & Schmidt, 1992)	Medicine	C	Paper case: pancreatitis
(Corcoran, 1986a, 1986b)	Nursing	C	Paper cases pain control in hospice setting
(Corcoran, Narayan, & Moreland, 1988)	Nursing	C	Telephone triage cases
(Elstein et al., 1978)	Medicine	R	Simulated Patients
(Elstein, Loupe, & Erdmann, 1971)	Medicine	R	Simulated Patients
(Embrey, Guthrie, White, & Dietz, 1996)	Physical Therapy	R	Real Patients
(Fonteyn & Fisher, 1995)	Nursing	C	Real Patients
(Gale, 1982)	Medicine	R	Real Patients
(Gale & Marsden, 1982)			
(Greenwood & King, 1995)	Nursing	C / R	Real Patients
(Henry et al., 1989)	Nursing	C / R	Computer simulation of a nursing problem
(Isles, 1995)	Physiotherapy	R	Real Patient in Neurology
(Jennett et al., 1992)	Medicine	R	Stimulated recall using the patient chart notes after the examination as the prompt
(Joseph & Patel, 1990)	Medicine	C	Paper case in endocrinology
(Neufeld, Norman, Feightner, & Barrows, 1981)	Medicine	R	Four Simulated Patients
(Payton, 1985)	Physical Therapy	R	Real patients
(Rivett & Higgs, 1997)	Physical Therapy	C	Videotape Observation of History Taking encounter
(Traband & Dunn, 1988)	Respiratory Therapy	C	Computer simulation of case

4.11 Methods for the Analysis of Qualitative Data or Verbal Records

In qualitative research, particularly in education and the health sciences, conceptual frameworks are often tested out in the field (Miles & Huberman, 1994; Vockell & Asher, 1995). This is carried out by observing phenomenon in the field and attempting to gain a holistic overview of the context and perceptions of the subjects. By capturing this data, which is generally in text form, certain themes can be identified which can be used to explicate the ways subjects account for, understand and take action. Qualitative methods also have a strong potential for supplementing, validating, explaining, illuminating or re-interpreting quantitative data gathered from the same setting (Miles & Huberman, 1994; Vockell & Asher, 1995).

Several investigators who have studied the CR processes of health professionals describe various methods for the analysis of verbal record data (Barrows et al., 1982; Bordage & Lemieux, 1986; Edwards, Jones, Carr, & Jensen, 1998; Embrey et al., 1996; Gale & Marsden, 1982; Greenwood & King, 1995; Jensen et al., 1992; Jensen, Shepard, & Hack, 1990; Miles & Huberman, 1994; Payton, 1985; Rivett & Higgs, 1997; Thomas-Edding, 1987). These qualitative methods, which are described below, employ techniques which allow inferences to be made about thinking processes and systematically and objectively, identify, categorise and classify the content of the subjects' accounts (Gale & Marsden, 1982; Greenwood & King, 1995; Jones, 1989).

4.11.1 Multiple Cases

The use of multiple or several cases provide the researcher with a deeper understanding of processes and outcomes as one can actually test hypotheses and gain a good picture of causality (Miles & Huberman, 1994). Multiple cases may also boost generalisability, although this goal is inappropriate for a qualitative study (Denzin, 1983; Guba & Lincoln, 1981).

4.11.2 Coding

In order to analyse textual data, which can be quite enormous, data reducing strategies are needed to gain a holistic understanding of the information contained in the written passages (Miles & Huberman, 1994). This is accomplished by coding the

data. Codes are tags or labels for assigning units of meaning to the descriptive or inferential information compiled during a study. They are used to sort and sift through text in order to identify patterns, themes, processes, commonalities and differences in sub-groups. The codes are usually attached to chunks of data of varying size. For example: words; phrases; sentences; or whole paragraphs. These codes are used to retrieve and organise these chunks of information into clusters and matrices which allow the investigator to draw conclusions and test out theories.

Codes can be descriptive, interpretive or used to denote patterns (Miles & Huberman, 1994). Descriptive codes describe a class of phenomenon in the text, whereas interpretive codes have a meaning that has been ascribed by the investigator. Both of these codes are first level codes and are useful for summarising data. Codes which denote patterns are highly inferential and explanatory and illustrate an emerging theme that has surfaced across the texts. These are second level codes and help to reduce data, particularly in cross-case analysis, so the researcher can begin to analyse the text, develop cognitive maps and denote emerging themes.

Codes to define or describe the text can be developed in tandem with the analysis of the data. This approach helps to create codes that are nested in the context of the data and is called the grounded approach (Guba & Lincoln, 1981). Strauss (1987) cited in Strauss and Corbin (1990) describe this process in further detail and advocate reviewing text line by line and indicating either beside or below each paragraph categories or labels that define the meaning of the text. These labels can then be reviewed and made more abstract so they can be attributed to several incidents. Material can then be categorised and differentiated to give you a sense of the frequency of each code.

This grounded approach suggests that codes are not static, but change throughout the analysis of the text. Lincoln and Guba (1985) give this dynamic coding strategy four labels: filling in; extension; bridging; and surfacing. Filling in involves adding codes or reconstructing schemes as new insights emerge. Extension involves returning to already coded text and interrogating them in a new way. Bridging involves recognising new or previously not understood relationships within a category and

constructing, and possibly creating, a new category. Finally, surfacing involves identifying a new category.

The health sciences literature is rich with examples of coding techniques for the study of clinical reasoning (Bordage & Lemieux, 1986; Christensen, 1993; Edwards et al., 1998; Embrey et al., 1996; Gale, 1982; Jensen et al., 1992; Jensen et al., 1990; Jones, 1995; Patel & Groen, 1986a; Patel & Groen, 1986b; Payton, 1985; Rivett & Higgs, 1997; Thomas-Edding, 1987; Whelan, 1988). Many of these codes have been used to guide the conceptual coding process in this study. Bordage and Lemieux (1986), for example, outline a very interesting coding system for studying the CR process of medical students. They point out that the techniques used by Elstein et al. (1978) led to a macroscopic description of clinical reasoning. It failed, however, to delineate differences in expert and novice reasoning. In light of this, they developed a series of codes to evaluate the knowledge structure of medical students engaged in clinical reasoning. The coding framework is more microscopic in nature and analyses the mental operations and knowledge organisation of the students by examining the structural semantics of the verbal records. Jones (1995) and Rivett and Higgs (1997) also describe several categories of CR in physiotherapy, which were described earlier. They are also useful for the development of coding schemes.

4.11.3 Coding Accuracy and Reliability

Training another person to analyse the textual data using the same codes is recommended (Corcoran, 1986a; Corcoran, 1986b; Elstein et al., 1978; Embrey et al., 1996; Fonteyn et al., 1993; Gale, 1982; Jensen et al., 1990; Miles & Huberman, 1994; Swanson, 1990). This is particularly important when one is using definitions/themes/categories from other sources. For inter-rater reliability between the two judges the percent agreement of coding and the Kappa coefficient (Cohen, 1960; Kazdin, 1982) are applied. These procedures have been used by others to monitor rating reliability (Corcoran, 1986a; Embrey et al., 1996; Jensen et al., 1990). Intra-rater agreement can also be carried out by having the principal coder code the same transcript three days apart (Embrey et al., 1996) and applying the same statistical tests. Coding reliability in the order of 90 per cent accuracy is

recommended and five to ten pages of text should be used for measuring coding reliability (Miles & Huberman, 1994).

4.12 Characteristics and Limitations of Qualitative Data Analysis

It is important to note that there are several limitations associated with qualitative data analysis. With respect to the process of transcribing the verbal text, accuracy is important. Accuracy is influenced by several factors. First and foremost, a transcriptionist with knowledge of the language and jargon common to the field is a bonus as it minimises transcription errors. The amount of detail in the transcription may also vary and needs to be specified. For example, are ‘umms’, ‘ahhs’ and pauses included/noted in the text. Many of the non-verbal aspects of communication are also lost when transforming verbal information to text.

Qualitative data is also not as generalisable as quantitative data as the latter usually emerges from large samples (Barrows et al., 1982). However, the goal of qualitative research is to give detailed descriptions of phenomena, not to determine generality (Patel & Arocha, 1995). The subjective elements of text interpretation also raise problems associated with reliability and reproducibility (Elstein, Shulman, & Sprafka, 1990). It is not uncommon for investigators engaged in a qualitative review of their data to: overweight facts they believe in or depend on; ignore or forget about data not going in the direction of preferred reasoning; and to see confirming instances far more easily than disconfirming instances (Nisbett & Ross, 1980; Vockell & Asher, 1995). The context in which the data is acquired also has important influences on what is verbalised and what can be concluded (Elstein et al., 1990).

Many of the experimenter biases can be reduced. Miles & Huberman (1994) describe 13 tactics that can be used to extract meaning from the data along with 13 tactics for testing and confirming these findings. These tactics are designed to minimise the multiple sources of analytical bias that can weaken or invalidate findings. Some example tactics for extracting meaning are: noting patterns or themes; clustering; counting; and making contrasts/comparisons. Examples of tactics that can be used for testing and confirming findings include: checking for representativeness;

triangulation or using multiple modes of evidence; weighting the evidence; checking outliers; and replicating a finding.

The use of verbal records to evaluate CR has been the focus of this remaining section of the chapter. It has highlighted the various methods that can be used to study the thought processes of health professionals and students. Further, this section has attempted to illustrate both the advantages and disadvantages of this study method. Ideas for coding and describing data have also been described as many of these methods will be used to evaluate the CR of the students in this study.

Chapter 5: Conceptual Framework

The review of the literature that has been presented thus far provides the background for the conceptual framework for this study. Miles and Huberman (1994) define a conceptual framework as a model which “explains either graphically or in narrative form, the main things to be studied - key factors, constructs or variables - and the presumed relationships among them.” (p.18). Figure 5.1 illustrates this framework .

The model that is presented builds upon the work of Elstein et al. (1978) and his hypothetico-deductive model of clinical reasoning. To illustrate the framework for this study, the clinical reasoning (CR) models developed by Barrows (1990) and Higgs & Jones (1995), illustrated previously as Figures 4.2 and 4.3, have been incorporated into a cone. This cone has been divided into three sections representing the cognitive (knowledge), psychomotor (skills), and affective (attitudes) domains of clinical competence. These three components are cited in the literature as being the central elements for task competence (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956; Corsini, 1984; Shepard & Jensen, 1999). As can be seen in the diagram, the CR spiral in the reciprocal peer coaching (RPC) model is contained within a much larger cone in comparison to the individualistic (IND) model. In many ways the broader RPC spiral is analogous to a much more powerful tornado, than the IND spiral. This wider RPC conic volume represents the greater breadth of learning, clinical reasoning and metacognitive activity that is likely to occur in the RPC model through the sharing of knowledge, skills and behaviours between peers. This greater breadth emanates from the peer based discussion and coaching which, in turn, influences the constructivist learning aspects of the patient encounter. Metacognition is heightened and the critical cognitive conflicts that emanate deepen the learning and clinical reasoning that eventuates. The theoretical basis for this enhanced learning and reasoning outcome in the RPC spiral is grounded in the discussion contained within sections 4.8.2 through to 4.9.2.

From within this conceptual framework the researcher has developed a hypothesis that peer assisted learning, or in this case the RPC model, increases the knowledge, cognitive, and metacognitive dimensions of the clinical learning and reasoning

experience. Through this increase, individuals ascend the clinical reasoning spiral with greater power and depth leading to enhanced performance in the cognitive, psychomotor and affective domains of clinical competence. The research methods used to validate this conceptual framework are described in the next chapter.

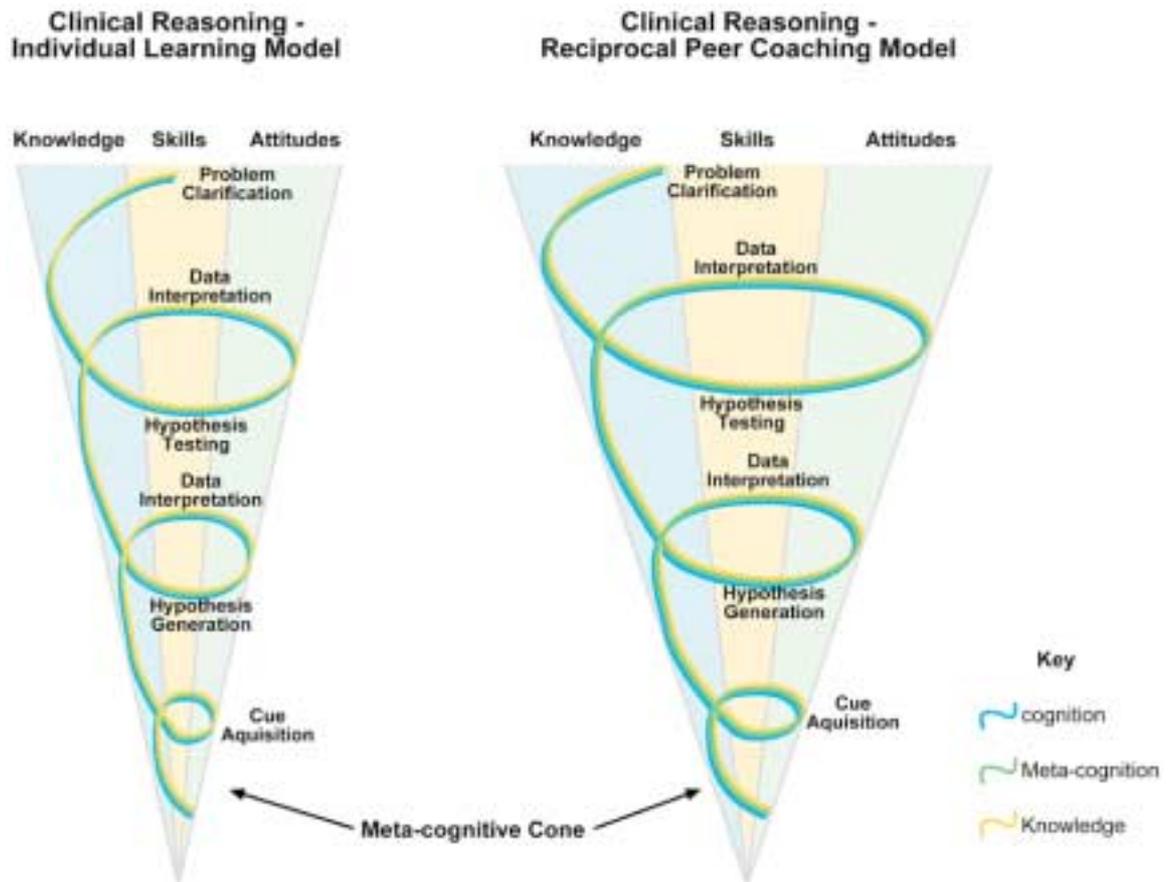


Figure 5.1: Conceptual Framework

Chapter 6: Methods

6.1 Research Design

This chapter outlines the research methods used in this study of reciprocal peer coaching. The study received ethics approval from the Curtin University of Technology Graduate Studies and Research Ethics Committee as part of the doctoral candidacy approval process.

6.1.1 The Simulated Patient

A single simulated patient (SP) case was used in this study. The SP case was a complete simulation incorporating history, physical and affective components. The use of complete patient simulations have been used by others to investigate clinical reasoning (CR) (Connell et al. 1993; Day, Hewson, Kindy, & Van Kirk, 1993; Ferrell, 1995; Rosebraugh et al., 1996). A SP case was selected as it allowed the investigator to control the patient factor (Helfer, Black, & Teitelbaum, 1975; Tamblyn et al. 1988; Williams et al., 1987). This control was needed to effectively measure differences in performance between students in the individual (IND) and reciprocal peer coaching (RPC) groups.

6.1.2 Case Selection for this Study

The students' academic and clinical background were considered in selecting an appropriate case for this study. A rotator cuff problem, which is commonly encountered in clinical practice, was chosen. An obscure case was avoided as it could lead to artificially low results and make it harder to discriminate between the two groups (Elstein, Shulman, & Sprafka, 1978; Newble, van der Vleuten, & Norman, 1995; Stillman, 1993).

The pathophysiology and management of rotator cuff problems are well addressed in the undergraduate physiotherapy curriculum at Curtin University of Technology. Rotator cuff problems are also common in clinical practice. Dekkar, van Baar, Curfs, & Kerssens (1993) studied 74 Dutch physiotherapists in 30 private practices. Over 8,700 patient episodes were analysed over three years. Out of the top ten conditions

seen by these clinicians, shoulder symptoms/complaints were ranked sixth. Rotator cuff pathology has also been simulated safely and successfully by other academic programs thus making it a suitable choice for this study (Brook, 1997; Gliva, 1997) .

6.1.3 Development of the SP Case

The patient that was used to develop the simulation was identified by contacting physiotherapists in the Perth metropolitan area. A patient with a rotator cuff problem was identified and invited to serve as the model for this study. Two people were asked to assess the patient. One was a physiotherapist with a Post-Graduate Diploma in Sports Physiotherapy. The other was a third year physiotherapy student, in good standing, who was at the same point in the undergraduate curriculum as the intended subjects for this study in the following year. These two assessments provided the investigator with a broader perspective of how a physiotherapist and student approached the patient evaluation.

The patient, physiotherapist and student received explanations about their role in the study and were asked to sign a consent form. The assessments were then videotaped by the investigator using a Panasonic M40 videocamera. Following the videotaping, the physiotherapist and student were asked to describe their principal and differential diagnoses and management plans for this client.

The SP case was then developed using the following sources of information: videotape recordings of the two above mentioned encounters with the real patient; supplemental information from the literature on orthopaedics and rotator cuff pathology (Adams & Hamblen, 1990; Caillet, 1981; Hartley, 1990; Jones, 1997; Maitland, 1977; Palmer & Epler, 1990; Paris, 1985; Potter, 1995); the investigator's ten years of clinical experience in outpatient orthopaedics/sports physiotherapy; and input from a panel of four physiotherapists and one student. A review of the unit outlines and instructional manuals was also undertaken to ensure that all potential activities and tests that students may employ were incorporated into the patient training notes and evaluation instruments (Pickard, Potter, Boyle, & Owens, 1996; Potter, 1995; Potter, Pickard, Williams, & Liston, 1996; School of Physiotherapy, 1996; School of Physiotherapy, 1997).

6.1.4 Developing the Measurement Instruments for the SP Case

A series of evaluation instruments were developed taking into consideration many of the design principles described in Chapter 3 (Colliver & Williams, 1993; Dawson-Saunders, Verhulst, Marcy, & Steward, 1987; McClure, Gall, Meredith, Gooden, & Boyer, 1985; Stillman, Ruggill, Rutatla, & Sabers, 1980; Traub & Rowley, 1991; Vu et al., 1992). The specific evaluation instruments that were developed were: a history checklist (HC), a physical examination checklist (PEC), a modified version of the Arizona Clinical Interviewing Rating Scale (ACIRS) and eight post-encounter questions (PEQ) designed to evaluate the clinical reasoning of each student involved in the study.

History and Physical Examination Checklists

Checklists were developed from the information that was contained within the SP case. To boost the discrimination of the checklists, each checklist item, where possible, was designed to describe one general skill at a time. The checklists were also written with minimal medical jargon. The scoring system for the HC and PEC used a three point rating system: zero (0): indicating ‘not done’; one (1): indicating ‘done poorly/incompletely’; and two (2): indicating ‘done well’. A three pronged scoring approach was considered to be more appropriate than a dichotomous scoring approach as the former leads to better discrimination among students (Page & Bordage, 1995; Page, Bordage, & Allen, 1995; Rethans, Sturmans et al. 1991; Ripkey, Case, & Swanson, 1996; Rothman & Cohen, 1995; Sloan et al. 1994). It also enabled better judgments to be made about the performance of the students. For example, in order to receive a score of two for the item dealing with active range of motion testing, a subject would have to mention that flexion, abduction, and external rotation needed to be tested. A response with less than these three tests would receive a lesser score.

The Arizona Clinical Interviewing Rating Scale (ACIRS)

The ACIRS is an interviewing and interpersonal communication skills rating scale that has been developed for use in clinical practice (Stillman, 1980; Stillman, Brown, Redfield, & Sabers, 1977). This scale was modified by the investigator for use in this particular experiment. Only 12 of the original 14 items were selected. The two items that were eliminated were not considered relevant for this particular study by the

investigator. One item focused on the use of transitional statements whereas the other item focused on summarisation. These two items were structured to support a medical style of interview where a review of organic systems normally takes place. Since physiotherapists do not conduct such a system review as part of their interview these two items were removed from the scale.

Each interviewing item on the original ACIRS is scored using three verbal anchors and a five point rating scale. This format allows raters to score a candidate between two verbal anchors. For this study, the three verbal anchors were maintained but a three point rating scale was used instead. This required the rater to select a verbal anchor that described the candidate's communication performance. A halfway between response was not permitted. This change helped to improve the inter-rater reliability of the ACIRS ratings. It was also consistent with the three point rating scale used in the other two checklists.

Post-Encounter Questions (PEQ)

Eight questions were developed for the post-encounter questionnaire. The first three questions were open-ended in format. The last five questions were a modified version of the multiple choice format. The first two questions asked students to describe their diagnoses and differential diagnoses for the SP encounter. These questions explore the students' hypotheses in the area of neuro-musculoskeletal physical diagnosis. The third open-ended question required students to delineate their management plans for the client.

Students were given a score of eight for the correct diagnoses (four for identifying the rotator cuff lesion and four for identifying the bicipital tendinitis lesion). A maximum score of ten was possible for the differential diagnosis question. Students were advised that they would be penalised one mark for every incorrect differential diagnosis listed as an answer. This penalty was incorporated to ensure that the students critically analysed their findings in coming to an answer. It was also put into place to minimise the possibility of a test taking strategy where students write down all possible answers to a question whether they are correct or incorrect. A maximum score of 20 was given for question three which required students to outline their management plans for the client.

Questions four through to eight were designed using the “extended matching format” (Case & Swanson, 1993; Solomon, Speer, Perkowski, & DiPette, 1994), “pick ‘n’ format” (Ripkey et al., 1996) and “key features” format (Page & Bordage, 1995; Page et al., 1995) described in the literature. All of these formats are similar and can be used to evaluate diagnostic skill, therapeutic skill or understanding of pathophysiology. Each question contained a total of 15 items with two ‘distracters’ for each correct key feature. Students were advised that they had to select five key features. A three point scoring system was selected for evaluating the student’s ability to correctly select the key features of the history and physical examination. The three point scoring system worked in the following manner: zero (0) was awarded if an incorrect distracter was selected; one (1) was awarded if a piece of information from the history or physical examination, which was perhaps salient but not a key feature, was selected; and two (2) was awarded if the correct key feature was selected.

6.2 Face and Content Validity of the SP Case and Test Materials

The panel of clinicians used in this study to establish face/content validity consisted of the following members: (1) a faculty member with experience teaching orthopaedics to both undergraduate and post-graduate physiotherapy students; (2) a faculty member with experience teaching the manual therapy component of the orthopaedic units to the undergraduate physiotherapy students; (3) a senior physiotherapist in a hospital-based outpatient physiotherapy service; (4) a Curtin University Clinical Tutor with experience supervising novice physiotherapy students in outpatient orthopaedic settings; and (5) a pre-clinical physiotherapy student who had just completed his third year of academic study. Table 6.1 outlines the experience of the first four panel members in more detail.

Table 6.1: Panel Members

Panel Member	Actual Years of Clinical Experience	Actual Years of Clinical Experience in Orthopaedics	Actual Number of Years Teaching/Supervising Physiotherapy Students
1	14.0	12.0	6.0
2	14.0	9.0	3.5
3	16.0	13.0	10.0
4	12.0	11.0	1.0
5	Fifth panel member is the third year undergraduate physiotherapy student		

This panel of clinicians reviewed all of the evaluation instruments and the SP case. This review required two meetings of two hours duration each. At least 80 per cent agreement on study criteria was considered an acceptable benchmark. Further, the panel were asked to specify whether a checklist item was essential or non-essential to minimise excessively high scoring standards.

6.2.1 Review of the History and Physical Examination Checklists

Panel members were initially shown the videotaped encounter between the student physiotherapist and the real patient. They were asked to make notes about the history and physical examination sections of the evaluation that they felt were essential, important or non-essential. Further, they were asked to identify any essential questions or procedures that they felt were omitted. Essential items were defined as necessary parts of the evaluation which yield information that would direct the clinician to the correct diagnosis/management. Important items were defined as necessary parts of the evaluation (thoroughness), but not necessarily yielding information that would direct the clinician to the correct diagnosis/management. Non-essential items were seen to be superfluous or contributing very little towards correct diagnosis/management.

The panel reviewed the draft HC and PEC, which was designed by the investigator. Throughout the process, the panel were reminded to keep in mind the competency level of the students in this study. An attempt to achieve a balance between thoroughness and citing only the essential items was part of the review process. The panel members were asked whether they agreed/disagreed with each item on the HC and PEC with respect to it being essential, important or non-essential. Further, any

items which they felt were essential/important but were omitted from the list were incorporated into the checklists. Only those essential/important items that had at least 80 per cent agreement were included in the checklists. Items considered to be non-essential, were eliminated in the same manner. Modifications to the nomenclature used in the checklists and descriptions of the expected behaviour/standards for each checklist item were also solicited from the panel. Copies of the final history and physical examination checklists, with accompanying interpretive guidelines, are attached as Appendix 1 and 2 respectively.

6.2.2 Review of the Arizona Clinical Interviewing Rating Scale

The ACIRS, as modified by the investigator, was also shown to the panel. They were asked to note whether each item was relevant for a novice physiotherapist conducting an interview. Again, at least 80 per cent agreement was needed for each item. A copy of the ACIRS is attached as Appendix 3.

6.2.3 Post-Encounter Questionnaire Review

The panel were instructed to review the PEQ in a manner similar to the other instruments. For the first three open-ended questions, panel members were required to agree/disagree with a list of diagnostic, differential diagnostic and management possibilities for the case that were developed by the investigator. They were also encouraged to suggest other possibilities that may have been omitted. Eighty (80) per cent agreement, or more, was still used as the benchmark for accepting an answer as a possible diagnosis, differential diagnosis or management option.

The panel were then instructed to review the two history and three physical examination key feature questions. They were given a brief explanation of the nature and purpose of this question format as it was unfamiliar to the panel. Each item within the menu of options was reviewed by the panel to determine whether it was a key feature of the case, important but not a key feature, or an appropriate distracter. Any key history and physical examination items that were not on the list but seen to be essential were also noted. Eighty (80) per cent agreement was again used as the benchmark for accepting/rejecting an item. Throughout the review process, the panel were asked to comment on the terminology used in the questions and to suggest

alternatives. A copy of the PEQ, with accompanying answers, is attached as Appendix 4.

6.2.4 Review of the Simulated Patient Training Notes

The final component of the face/content validity review session required the panel to review the training notes for the simulated patient. Panel members were asked to read through the notes and to highlight any concerns they may have about the nature of the prescribed SP response or the types of questions and tests that were going to be used in the case. Panel members were again reminded that the station should be reviewed from the perspective of a third year physiotherapy student with minimal clinical experience. Further, any questions or physical examination maneuvers that were absent from the training protocol, which were considered essential by the panel members, were noted and incorporated into the training notes. Any controversies, such as whether an item was considered an entry level competency, were settled by ensuring there was at least 80 per cent agreement among the group. A copy of the training notes used to guide the training of the SP is attached as Appendix 5.

6.3 Simulated Patient Recruitment and Training

6.3.1 Recruitment and Selection Issues

For this study, a male actor was recruited from a company that employs professional actors for training and development purposes. The age and physical characteristics of the SP case were used as guidelines for the selection of an appropriate actor. The same actor was used for both the pilot and main study.

6.3.2 Simulated Patient Training

The training of the SP was carried out using procedures described by Barrows (1987) and the information provided in Chapter 3. The training procedures for the SP in this study are summarised below. First, the SP was given an opportunity to observe the real patient on videotape. Second, aspects of the SP's own personal history were incorporated into the case to increase the realism and spontaneity of the presentation. Training procedures avoided the use of medical jargon to keep the SP appropriately naive (Barrows, 1987; Stillman, 1990). The SP did not receive a copy of the training

notes. The notes were used to direct the investigator, who also served as the patient trainer, to prepare the actor. The actor was encouraged to keep his own notes so he would come to 'know' the case from his own perspective.

6.3.3 Practice Sessions

Once the investigator felt that the SP could portray the case adequately, five fourth year physiotherapy students and two physiotherapists volunteered to evaluate the patient. All members signed informed consent and confidentiality forms. Of the five students selected, two had completed a fourth year placement in outpatient orthopaedics whereas the other three had yet to gain experience in that area. Two of these five students evaluated the SP as a pair. Both high and low achievement levels among the students, using their composite scores in the undergraduate orthopaedic units, were represented. The two physiotherapists had at least two years of clinical experience, including at least six months of practice in outpatient orthopaedics.

After each of the physiotherapists and students had an opportunity to evaluate the SP, they were asked to comment on the face validity of the patient and to note any patient responses that did not make sense, were inconsistent, or unrealistic. The encounters were also videotaped and used to illustrate key training points to the SP. The SP also rated the performance of the students and physiotherapists using the HC, PEC and the Arizona Clinical Interviewing Rating Scale. The ratings of the SP during these training sessions were compared to the ratings of the investigator, who observed the encounters directly. Discrepancies in ratings were used to clarify training procedures and to monitor the reliability of the simulated patient. Once the investigator was comfortable with the performance accuracy and rating reliability of the SP, a pilot test was set up to ensure the main experiment would run smoothly. A total of 30 hours was spent training the SP to a level appropriate for pilot testing.

6.4 The Pilot Test

A formal pilot test was initiated to further test the fidelity, reliability and accuracy of the SP over a number of IND and RPC encounters. The pilot test was also used to determine: the average time spent in the patient encounter by the students; how well students understood and completed the various questionnaires; the construct validity

of the measurement instruments; and whether the station standards set by the panel of experts were relevant and achievable.

6.4.1 Subjects

Subjects were recruited from the School of Physiotherapy and the clinical community. Twelve fourth year physiotherapy students with minimal or negligible outpatient orthopaedic experience were selected for the study. Four physiotherapists with at least two years of experience, including at least six months of experience in outpatient orthopaedics, were also recruited from the clinical community. Four of the students evaluated the SP on an individual basis. The remaining eight students were organised into RPC pairs. These pairs were responsible for evaluating the simulated patient. Both low and high achieving students, using academic achievement scores in their undergraduate orthopaedic units as the benchmark, were used in the pilot test. These students were spread across the IND and RPC groups. The four physiotherapists assessed the patient independently. Hence, the pilot test consisted of 12 SP encounters (4 IND-student, 4 RPC-student, 4 Physiotherapist) which took place in an environment that resembled a clinical setting. All encounters with the SP were videotaped. Informed consent and a statement of confidentiality were also obtained from all participants in the pilot test.

6.4.2 The Simulated Patient Encounter

Prior to the SP encounter, each participant completed a confidence/anxiety pre-encounter questionnaire. This questionnaire is attached as Appendix 6. Students and physiotherapists were instructed to carry out a 'comprehensive assessment'. This overt explanation was given to minimise the physiotherapists from relying solely on pattern recognition to work through the case and to demonstrate their competence.

Participants were all provided with a standard outpatient assessment form which highlights the main categories of inquiry that need to be performed. When the participants finished the history section of the interview, they were required to leave the assessment area to plan their physical examination. This five minute interval was used by the SP to complete the history checklist. Upon the completion of the physical examination, the SP completed the PEC and the Arizona clinical

interviewing rating scale. In the case of the RPC group, the HC, PEC, and ACIRS was scored on a collective basis - i.e., the dyad received a single score for their combined performance.

6.4.3 Post-Encounter Activities

Following the SP encounter, the participants were asked to complete a confidence/anxiety post-encounter questionnaire and to rate the fidelity of the simulation. A copy of the questionnaire can be found in Appendix 7. The fidelity ratings are provided in Table 6.2.

Table 6.2: Fidelity of the Simulated Patient

Group	Overall Score		History Section		Phys. Exam. Section		Psychosocial Aspects	
	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
Students	4.41	.79	4.08	.79	4.33	.65	4.25	.87
PTs	4.25	.50	4.50	.58	4.00	.82	4.25	.96

n = 12 students; n = 4 physiotherapists (PTs)

Measured using a 5 point Likert scale - 1=unrealistic, 5=very realistic

Both the students and the physiotherapists found the simulation to be realistic. Physiotherapists found the SP's physical examination slightly less realistic. The SP was trained to a level which would be appropriate for third year undergraduate physiotherapy students. Occasionally the physiotherapists would apply advanced techniques to the simulated patient. Given that the SP had not been trained to respond to these advanced approaches, the SP would occasionally provide the incorrect response. In spite of these very infrequent occurrences, the physiotherapists still found the simulation to be realistic, in particular, the history section.

After the completion of the SP encounter, the participants were then given up to 15 minutes to think about the case. Members of the RPC were permitted to discuss the encounter with their colleague. After this 15 minute interval, each participant was required to independently answer the post-encounter questionnaire. The time required to complete the PEQ was collected for each individual in order to determine a time frame to complete the PEQ for the main study. The students, on average, completed

the PEQ in 19'30" with a range of 13'03" to 25'00". A maximum of 30 minutes for the PEQ was established for the main study in light of these results.

6.4.4 Determining Station Duration

The establishment of an appropriate station duration was needed to ensure that the students would be able to complete the history and physical examination and to minimise the possibility of 'first visit bias' (Tamblyn et al., 1992). First visit bias can have a negative impact on outcome scores if the case cannot be managed adequately in one interaction. Hence, practice sessions and the pilot study were used to determine a realistic time frame for completing the complete SP encounter.

None of the participants in the practice sessions and pilot study operated under a time control. It was important to find a station duration that did not advantage or disadvantage the groups. As a result, the mean time and range of times taken by both the IND and RPC groups to fully assess the SP in the practice sessions and pilot study were used as guidelines for setting the station duration for the main study. The mean time taken to complete the SP evaluation by the students was 29'46" with a range of 20'15" to 46'00". The mean time taken to complete the SP evaluation by the physiotherapists was 36'44" with a range of 26'25" to 47'30". Given the time needed for the SP to complete the PEC and ACIRS, the likelihood that the subjects for the main study will take longer given their novice status and the time frames obtained from the students and physiotherapists, a 60 minute station duration was selected for the main study.

6.4.5 Accuracy, Inter-Rater & Intra-Rater Reliability of the Simulated Patient

Accuracy is present when there is reproducibility in the SP presentation and is measured by recording the proportion of essential clinical features presented correctly in each SP encounter (Tamblyn, Klass, Schnabl, & Kopelow, 1991a; Tamblyn et al., 1988). Accuracy was measured by the investigator during the pilot study by observing the SP's performance directly and recording the proportion of clinical features presented correctly in each encounter. The listed items on the HC and PEC were used to track this performance accuracy. The SP proved to be very accurate.

History findings were presented accurately 96.95 per cent of the time. Physical examination findings were presented accurately 94.70 per cent of the time.

With respect to inter-rater reliability, the ratings of the SP on the HC, PEC and ACIRS during the practice sessions and pilot study were compared to the ratings of the investigator. The investigator has ten years of experience in musculoskeletal practice and developed the instrumentation and training protocol. Hence, he was able to serve as an expert rater and was a suitable criterion reference point.

Percent agreement and the Kappa coefficient were used to evaluate rating reliability. These procedures have been used in other studies to evaluate rater agreement (Corcoran, 1986; Embrey, Guthrie, White, & Dietz, 1996; Jensen, Shepard, & Hack, 1990; Vu et al., 1992; Williams et al., 1987). The Kappa coefficient is a useful measure for evaluating inter-rater reliability as it provides an estimate of agreement between observers corrected for chance (Cohen, 1960; Kazdin, 1982). Factors which can influence rater reliability are observer bias, observer drift and expectancy bias (Kazdin, 1982). Observer bias occurs when raters change their rating behaviour when they know they are being observed. Observer drift occurs when raters, over time, drift away from the original standards for evaluation. Expectancy bias occurs in raters by rating a specific group whom they expect should be doing better. The latter two forms of bias were minimised in this study by regular reviews of the SP's rating and ensuring that the SP used the written standards for evaluation. Continuous assessment of the rater to minimise observer bias was not possible so there may have been some bias during the observation sessions.

The investigator chose to unobtrusively observe the interactions rather than observe them later on videotape. Unobtrusive monitoring involved positioning oneself at the entrance to the treatment cubicle and observing the encounter. Videotape review can reduce inter-rater reliability measures because of the difficulty in capturing all information on film (Connell et al., 1993; Jain et al., 1997; Liu et al., 1980; Stillman et al., 1991; van der Vleuten & Swanson, 1990).

Intra-rater reliability of the SP was determined using a procedure described in the literature (McClure et al., 1985; Stillman et al., 1977). The SP observed six of the

student encounters that took place during the pilot study, on videotape, two weeks later and was asked to complete another series of checklists (HC, PEC and ACIRS). The percent agreement and Kappa coefficient between the original SP ratings and the SP ratings two weeks later were calculated as a measure of the SP's intra-rater reliability. The inter-rater and intra-rater reliability measures for the SP are outlined in Table 6.3.

Table 6.3: Inter-rater and Intra-rater Reliability of the Simulated Patient

Inter-rater Reliability		
Category	Per cent Agreement	Kappa Co-efficient
History	90.4	.85
Physical Examination	82.5	.79
History and Phys. Exam.	86.8	.87
ACIRS	82.4	.73
All Categories	85.9	.83
Intra-rater Reliability		
Category	Per cent Agreement	Kappa Co-efficient
History	91.0	.84
Physical Examination	77.5	.63
History and Phys. Exam.	83.8	.74
ACIRS	77.7	.55
All Categories	83.3	.72

Inter-rater reliability: 2 raters, 12 encounters

Intra-rater reliability: SP rater, 6 encounters

The inter-rater reliability indices were acceptable, being comparable to many of the values described in the literature (Colliver & Williams, 1993; Nayer, 1993; Norman, Muzzin, Williams, & Swanson, 1985; Swanson & Stillman, 1990; Vu & Barrows, 1994; Vu et al., 1992). Inter-rater reliability was generally lower on the ACIRS scale which deals with communication. Again, this is in keeping with trends seen in the literature (Bowman et al., 1992; Dawson-Saunders et al., 1987; Swanson & Norcini, 1989; Swanson & Stillman, 1990; Tamblyn, Klass, Schnabl, & Kopelow, 1991b). With respect to the inter-rater reliability measures, 13 out of the 37 rater disagreements (35 per cent) in the physical examination section were due to the investigator not seeing or experiencing tests that required palpation on the part of the

participant. Another 8 of the 37 rating disagreements (22 per cent) in the physical examination were due to multiple ways of conducting the same test which the SP was unable to rate. For example, five different ways of testing the bicipital tendon emerged from the original method used to train the simulated patient. The SP would not have recognised these alternative methods although the investigator did. Nonetheless, a rating discrepancy would have emerged. All rating discrepancies between the investigator and the SP were discussed and used to either clarify scoring procedures or to augment the SP's training.

The intra-rater reliability measures were generally lower than the inter-rater reliability measures although still within acceptable ranges. Tamblyn et al. (1991b) notes that intra-rater reliability estimates can be under-estimated because the videotape review process is contextually different from the original scoring episode. Further, Tamblyn et al. (1991a) note that 12 per cent of physical examination findings in one particular study could not be assessed because of the visual impediment created by the use of videotape. Simulated patients are also likely to be more critical when viewing the videotape because they do not have to rely on recall to complete the checklist, nor do they have to perform the tasks (Tamblyn et al., 1991b). Twenty-one of the 138 observations (15 per cent) that resulted in rater disagreements in this study were, in fact, due to problems of observation (i.e., the participant had blocked the camera view or the videotape was unable to capture the correct angle for visualisation).

6.4.6 Construct Validity of the Measurement Instruments

Gronlund (1981) notes that administering a test to known groups whose scores should differ as a result of training is one method of determining construct validity. This method was used in the pilot test by comparing the scores of the physiotherapists and students. This approach for measuring construct validity has been described in numerous studies that have employed SP based examinations (Newble, Hoare, & Elmslie, 1981; Norman, Feightner, & Tugwell, 1983; Petrusa et al., 1987; Robb & Rothman, 1985; Stillman et al., 1986). Table 6.4 illustrates the differences in scores between the students and the physiotherapists for the anxiety

and confidence questionnaires. Table 6.5 illustrates these differences in scores for the HC, PEC, ACIRS, and the post-encounter questionnaire.

Table 6.4: Independent Samples t-test for Mean Pre-Anxiety/Confidence

Item	Students n = 12		Physiotherapists n = 4		Effect Size	t value
	Mean	s.d.	Mean	s.d.		
Anxiety	3.33	0.9	2.75	0.5	-.70	.14
Confidence	2.75	1.1	3.75	1.0	.87	.13

Anxiety: 1 = no anxiety5 = high anxiety
Confidence: 1 = no confidence5 = high confidence

Table 6.5: Independent Samples t-test for Mean Scores on the Patient Encounter Checklist and Post-encounter Questionnaire

Item	Student Score n = 8		PT Score n = 4		Effect Size	t value
	Mean	s.d.	Mean	s.d.		
HC+PEC	57.8	7.3	70.7	9.1	1.39	.01*
ACIRS	86.2	2.0	88.4	1.0	0.41	.38
PEQ	57.3	5.6	65.4	4.9	1.27	.03*

*p<.05; scores expressed as a percentage

The anxiety and confidence scores in Table 6.4 demonstrate that the average student was more anxious and less confident than the average physiotherapist prior to the SP encounter. These differences, which have strong practical significance reflected by the effect size, were not statistically significant¹. This may be due to the small numbers of subjects in this pilot study. The anchoring points were also modified, in

¹ To see how effect sizes are calculated please refer to page 228-229.

light of the pilot study, to improve clarity. This was done because the anchoring points were at cross purposes. For example, a score of 'five' meant high confidence, but also high anxiety. For the main study, questions on anxiety and confidence were changed such that high anxiety and no confidence equaled a score of 'one' and no anxiety and high confidence equaled a score of 'five'.

Good construct validity was demonstrable for the history and physical examination checklists ($t = .01$; $df=10$; $p < .05$) and the post-encounter questionnaire ($t = .03$; $df=10$; $p < .05$) with the physiotherapists demonstrating statistically significant higher scores. There was no notable difference between the students and the physiotherapists on the Arizona clinical interviewing rating scale. This is not surprising considering the students in the pilot study were fourth year students who had already completed over 10 weeks of clinical practice. Hence, they would have had an opportunity to develop their communication skills. The SP was also a relatively cooperative patient. As a result, the need to implement high level communication and interpersonal skills was not necessary. Had the patient been difficult, one may have seen a greater distinction between the groups on communication skill.

6.4.7 Criterion Validity - concurrent

As was noted earlier in the literature review, one of the difficulties associated with measuring criterion validity in clinical performance evaluations is the lack of a gold standard to which comparisons can be made (Gronlund, 1981; Nayer, 1993; Newble et al., 1981; Stillman et al., 1977; Swanson & Stillman, 1990; van der Vleuten & Swanson, 1990; Vu & Barrows, 1994). In the case of this study, only academic grades for the undergraduate orthopaedic units were available for comparative purposes. These grades are comprised of knowledge tests and practical skill tests in orthopaedics. Given that the PEQ score is a clinical reasoning score in the area of orthopaedics, the PEQ scores were correlated with the combined orthopaedic unit scores for the students in the pilot test. The Pearson product moment correlation coefficient was calculated, using $r=0.57$. This suggests a fair/moderate positive correlation between the PEQ and the scores in the orthopaedic units.

The scores of the students in the pilot study that were related to history and physical examination were also correlated to the students' scores on the post-encounter questionnaire. The Pearson product moment correlation coefficient was applied. Performance on the history checklist was correlated with the key feature questions dealing with history on the post-encounter questionnaire. A correlation of $r=.39$ was found. This weak correlation is not a surprise given that the history questions asked by physiotherapists are fairly generic. Students who have determined the most likely physical diagnosis at the end of the encounter would still be able to guess some of the key history findings, even if they failed or forgot to ask them in the first place. For example, if a student is able to determine that the patient has a supraspinatus lesion through their physical examination, he/she could extrapolate that the patient is likely to be unable to sleep on the affected side; a key history feature. The lack of variance in the subjects' scores also tends to lower the value of the correlation co-efficient.

A fairly strong correlation ($r=.64$) was found between the PEC score and the key feature questions dealing with physical findings. This suggests that a good physical examination is positively related to success in completing this component of the questionnaire. A weak correlation was found between the combined history and physical examination checklist scores and the question dealing with diagnosis ($r=.25$). This is not surprising as the HC and PEC scores are measures of psychomotor skill. Commonly, students will ask correct questions and perform correct procedures but fail to interpret and/or synthesise this information correctly. The student, in answering the questions dealing with diagnosis may be using biomedical knowledge quite separately from the clinical information collected during the assessment.

6.4.8 Review of the Measurement Instruments

All items on the HC and PEC, and the key feature items on the PEQ, were evaluated to determine how the students responded to each of the items. This gave the investigator an indication of ambiguous or problematic items on the checklists, and whether items were far too easy or difficult for students. The item review was also a useful indicator for determining whether the expert panel had set the standards too high for third year students. The outcomes of this measurement instrument review

were then used to strengthen the quality of the measurement instruments for the main study. Table 6.6 summarises the results of the item analysis for the HC and physical examination checklist. Only those items where more than 50 per cent of the students failed to elicit findings are displayed.

A variety of explanations are offered for the results that are illustrated in Table 6.6. The HC items that appear to be problematic are considered to be part of a comprehensive or thorough history and certainly within the realms of novice competency. With respect to item 21, students may not be familiar with diagnostic ultrasound which may explain why they did not inquire about this piece of information. The remaining categories 23 - 25 reflect the poorer history taking abilities of the students and the fact that novices tend to focus on data gathering versus gaining a complete understanding the patient's illness experience (Jensen et al., 1990). The physiotherapists, in contrast, were much more successful in eliciting information pertaining to items 21, 23 and 24 by probing much further into the patient's illness history.

With respect to the PEC, items 8 - 10 were not carried out by any of the students. This may reflect a curricular gap, an exceedingly high expectation on the part of the expert panel who set the standards or something that the students may have done in a later examination (eg. first visit bias). Two of the four physiotherapists performed these examination items in full or in part. In spite of this poor showing, the items were retained on the examination for the main study as they may discriminate between high and low achieving students. To identify the possibility of a first visit bias, an open-ended question was added to the post-encounter questionnaire asking subjects to identify any elements of the overall assessment they chose not to perform and their accompanying reasons.

Table 6.6: Item Analysis for the History Checklist and Physical Examination Checklist

History Checklist

Item	Description	Per cent Failing to
------	-------------	---------------------

Elicit History Finding

21	Ultrasound Test and Significance of Test	100.0
23	Details of Physical Job Demands	62.5
24	Whether sport is played and frequency	75.0
25	Details of social situation: home; partner etc.	100.0

Physical Examination Checklist

Item	Description	Per cent Failing to Elicit Physical Examination Finding
8	Pull Shoulder Blades Together: sitting/standing	100.0
9	Pull Shoulder Blades Together: prone	100.0
10	Serratus Anterior Muscle Testing	100.0
11	Checks Neck Movements	62.5
21	Checks Elevation of Left Shoulder in supine (pec minor tightness)	100.0
22	Manipulates shoulder (tests accessory movements)	75.0
23	Accessory movement of Acromioclavicular joint	75.0

A possible reason for the students' failure to complete item 11 may stem from them ruling out this pathology during the history. However, clearing the neck in shoulder pathology is still a critical part of the examination so this item was not deleted from the checklist for the main study. Checking elevation of the shoulder for pectoral muscle tightness (item 21) also did not appear to be completed by a large number of the students. This test, however, can be completed fleetingly as part of the observation process. Hence, it is difficult to score. In light of this scoring difficulty, therefore, item 21 was removed from the physical examination checklist.

Interestingly, many of the students completed item 22 but did not carry it out according to the standards set by the expert panel. Hence, they did not obtain a score for this item. Again, this poor performance may be due to exceedingly high standards or a problem with how this skill was taught or understood by the students. The remaining item 23 was basically not completed by 75 per cent of the students. Again, students may have ruled out the possibility of acromioclavicular problems from their history and objective tests. Despite these errors, items 22 and 23 were retained on

the checklists as they may still facilitate discrimination between high and low achieving students.

Table 6.7 summarises the results of the item analysis for the key feature questions on the post-encounter questionnaire. Only those items where more than 50 per cent of the students failed to select the correct key feature are displayed.

Reasons for the students not selecting the key features outlined in Table 6.7 are varied. With respect to question 4.3, the students may not have known that Feldene is a non-steroidal anti-inflammatory drug and that the client's response to the drug was a key feature. Most students failed to select question 4.11 as a key feature even though it is one of the cardinal complaints of a patient with this particular syndrome. Most students selected the fact that the pain was superficial and sharp, or the fact that the patient reported no neck problems as alternatives. Even though these are relevant, they were not seen as key by the expert panel. In response to the poor selection of item 4.11, this item was reworded.

Table 6.7: Item Analysis for the Post-Encounter Questionnaire: Key Feature Questions 4 - 8

Item	Key Feature	Description	Per cent Failing to Select Correct Key Feature
4	3	Non-steroid Anti-Inflammatory Drugs Ease Pain	68.0
4	11	Moving Left Arm Forward is Painful	77.0
4	15	Positive Ultrasound Test	75.0
5	12	Unable to Sleep on Left Side	75.0
6	9	Tenderness Left Upper Trapezius	75.0
7	14	Grade 3 Caudad Accessory Movement	
		Glenohumeral Joint eases pain	92.0
8	5	Left Pectoral Minor Muscle Tightness	92.0

None of the students asked about diagnostic ultrasound in the history, yet, 25 per cent of the students selected question 4.15 as a key history feature. This illustrates guessing or cueing. The poor selection rate of question 5.12 reflects poor history technique and failure to probe adequately into the night time sleep patterns of the client. Again, most physiotherapists were able to ascertain this key feature.

The poor discernability of questions 6.9, 7.14, and 8.5 are tied to the poor technique of the students during the physical examination. Fifty per cent of the students failed to adequately palpate the upper fibers of trapezius and 75 per cent failed to properly evaluate the accessory movements of the left shoulder. Given this poor performance it is not surprising that they failed to answer items 6.9 and 7.14 properly. Item 8.5 was picked up by only one student despite the measurement difficulties associated with this item. Again, the students' failure to identify this item as a key feature may reflect a curriculum gap, an exceedingly high standard set by the expert panel or the possibility that this evaluation technique was not given much priority.

The item analysis, in the end, provided useful insights into the problem solving and CR skills of the students and facilitated further modification of the instruments. With respect to the SP, a couple of modifications were made to the PEC in order to boost his rating reliability. The first change was related to a two part test which evaluates the integrity of the supraspinatus tendon. Each part of the test was originally evaluated using a separate checklist item. However, most students combined the two

tests into one approach, making scoring difficult. Further, there were a variety of acceptable ways to conduct the test. Hence, these items were combined into one checklist item to ensure the SP could score them more objectively. The other modification related to the testing of the bicipital tendon. Again, there were several ways (5) of performing this test. These were categorised into one checklist item to ensure the SP could score these skills more objectively.

6.4.9 Qualitative Aspects of Clinical Reasoning

Methods to derive information about the qualitative aspects of CR were also employed as part of the pilot test. One of the students from the IND group and two students from the RPC group were selected from the group of 12 students. Each student was required to review their encounter with the SP on videotape. Before carrying out this review, the students practised retrospective recall using two exercises described in the literature (Ericsson & Simon, 1993).

After these two practice sessions, students watched their SP encounter and were instructed to verbalise their thoughts according to methods described in the literature (Barrows, 1987; Elstein et al., 1978; Embrey et al., 1996; Ericsson & Simon, 1993). This process was carried out by the investigator to become more familiar with this technique.

The students' verbal records were transcribed by a transcriptionist and then reviewed by the investigator by comparing them to the original audiotape. Any errors were corrected. A review of the transcripts by the investigator indicated that the verbal recall process could be improved. For example, the students tended to avoid stopping the videotape during the verbalisation process. This resulted in the students not completing some of their verbalisations as new cues would emerge from the ongoing interaction on the videotape and elicit another train of thought. Greater use of the pause button would minimise this problem. Students also became evaluative of their performance, which is a level three verbalisation. More prompting by the investigator, using prompts such as "what were you thinking" versus "keep on talking" should focus subjects on eliciting more level one and two verbalisations.

The verbal recall process also focussed on the CR aspects of the encounter. This did not provide much information on the RPC process itself. Hence, a series of open ended questions, at the end of the recall session, would be asked of subjects in the RPC groups. These questions would focus on the process of working in a pair. The open ended questions that were developed for the main study are described in Figure 6.1.

-
1. What did you like about the reciprocal peer coaching experience?
 2. What did not you like about the reciprocal peer coaching experience?
 3. Were there any aspects of the reciprocal peer coaching experience that facilitated your clinical reasoning or performance?
 4. Where there any aspects of the reciprocal peer coaching experience that hindered your clinical reasoning or performance?
 5. In what ways were you able to support your partner in preparation for the patient encounter?
 6. What suggestions can you make to improve the RPC process?
 7. If you could go back and repeat this experience, what might you have done differently?
 8. If you had a choice as part of your learning, would you prefer to always see patients individually, with a peer coach or as a combination of both? Why?
-

Figure 6.1: Reciprocal Peer Coaching Questions

6.5 Main Study

6.5.1 The Context

The subjects that participated in this study were third year undergraduate physiotherapy students (n=62) enrolled in a 15 week unit entitled, “Health And Social Behaviour in Physiotherapy (HSBIP) 350”. This unit has a lecture and tutorial component. The lecture component focusses on psychosocial issues related to health. The tutorial component is a communication unit that looks at the patient interview process, counseling and professional communication. This current study was incorporated into the context of the HSBIP unit as it had educational value and assured the investigator a large number of participants.

The nature of the students’ involvement was outlined at the beginning of the unit with comprehensive descriptions of their participation in their course manual. They were advised that the SP encounter would account for 25 per cent of their unit grade

and that this score would be posted in rank order, using student numbers, at the end of the unit. Students in the RPC group were advised that their final score would be based upon the dyad's actual HC, PEC and ACIRS score and the combined average of the dyad's individual scores on the post-encounter questionnaire. In other words, the combined approach of the dyad in carrying out the patient examination yielded a single HC, PEC and ACIRS Score. Dyads were given the freedom to plan their approach, the only condition that was imposed was that they each had to take responsibility for participating in either the history taking or physical examination. This ensured that both parties in the dyad played an active role in the patient assessment. For the post-encounter reasoning test, each member of the dyad carried out this test independently. However, their scores were averaged to yield a single post-encounter questionnaire score. This latter scoring strategy was put in place to ensure that students worked together cooperatively rather than competitively. To ensure that ethical standards for this experiment were not breached, all students were advised that if there was a significant difference in the mean scores for the IND and RPC group, the scores would be adjusted to ensure equity.

6.5.2 Sampling and Allocation to IND and RPC Groups

Students were allocated to the experimental and control groups by the investigator using their combined academic grades in orthopaedic science 252 and 351. Orthopaedic Science 252 is a one semester unit administered in the second semester of the second year program. It consists of the following modules: (1) orthopaedic lectures; (2) manipulative therapy - assessment and treatment of peripheral joints; (3) hydrotherapy; (4) functional rehabilitation; (5) assessment and treatment of musculoskeletal conditions; and (6) clinical management (Pickard et al., 1996; School of Physiotherapy, 1997).

Orthopaedic science 351 is also a one semester unit administered in the first semester of the third year program and consists of the following modules: (1) orthopaedics; (2) rheumatology; (3) amputees, burns and plastics; (4) manipulative therapy - assessment and treatment of the spine; (5) therapeutic exercise; and (6) hydrotherapy (Potter et al., 1996; School of Physiotherapy, 1996).

Both OS 252 and OS 351 have clinical visits as part of the unit. The former has three visits in an orthopaedic outpatient setting and all students learn in groups of six, but are often split up into smaller groups (triads/pairs) when working with patients. Students also complete a 10 day, full time clinical placement in a clinical setting during the second semester of second year. This is a physiotherapy assistant placement which is designed to help the students become familiar with a health care agency and professional communication. A variety of agencies are used for this placement experience. Orthopaedic Science 351 has eight clinical visits, two of which are in an outpatient orthopaedic setting. One visit is individualistic whereby the student visits a private practice. The second visit involves four students, although students are often split up into smaller groups (pairs) when working with patients. Hence, the students that participated in this study have minimal experience in orthopaedics even though they have completed the bulk of their studies in this area.

The 62 students were rank ordered from highest to lowest using their combined OS 252/351 grades. Students in the lower half of the distribution (53 - 71.5) were assigned to the 'low score' group while students in the upper half (71.5 - 90.5) were assigned to the 'high score' group. Figure 6.2 illustrates this distribution.

Students from the low and high score groups were then assigned to either the IND or RPC group. In order to evaluate the influence of IND and RPC models on performance outcomes, it was important to ensure that the academic achievement levels in both groups were similar. If by chance the two groups had significantly different academic achievement levels, the outcomes of the study could be due to academic differences alone. Table 6.7 describes the academic grades of the IND and RPC groups. Other control factors that influenced group assignment were age and gender. These two factors were controlled by balancing them across both groups. Table 6.8 provides an overview of the demographics of the IND and RPC groups. As is evident from the table, age and gender balance were equivalent across the two groups.

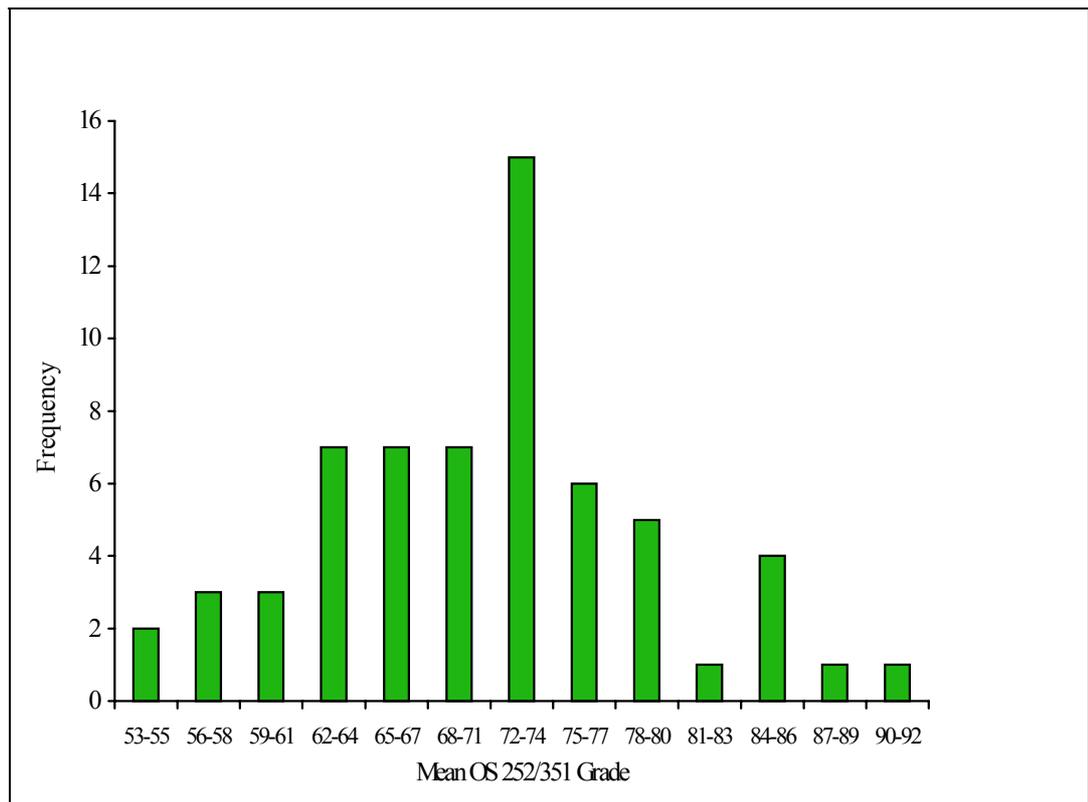


Figure 6.2: Distribution of Third Year Physiotherapy Students OS 252/351 Grades (n=62)

Table 6.8: Demographics of IND and RPC Group

Category	Variable	IND n = 20	RPC n = 42
OS Grade	mean	70.5	70.9
	s.d.	8.0	8.6
Gender	male	35 %	33 %
	female	65 %	67 %
Age	mode	19.0	20.0
	mean	22.0	22.2
	s.d.	5.0	3.6

The procedure for assigning students to the control and experimental groups used purposive or criterion based sampling (Jensen, Shepard, Gwyer, & Hack, 1992; Miles & Huberman, 1994). This sampling procedure assigns subjects to groups using certain characteristics or standards that are considered to be central units in the investigation. Samples are selected that fit the standards and are theoretically driven by a conceptual question rather than a concern for representativeness (Miles & Huberman, 1994). In this case, academic achievement level was the principle criterion for random assignment to a specific group.

The IND or control group (n=20) was composed of 10 low scorers and 10 high scorers. The RPC or experimental group (n=42 or 21 pairs) was composed of seven low/low scorer pairs, seven high/high scorer pairs and seven high/low scorer pairs. The high versus low distribution within the IND and RPC groups was implemented to see whether certain RPC combinations facilitated or hindered performance and reasoning. Evidence for this possibility can be extrapolated from work carried out by Bordage and Lemieux (1986). They demonstrated a difference in the frequency of mental operations in medical students considered to be strong or weak by their tutors.

Scores from ‘international’ students who were enrolled in the unit (n=2) were not included in this study. An international student for the purposes of this study was one whose first language was not English and/or they came from a country with a non-Western background. Issues related to culture can influence patient-clinician communication and practice and could possibly influence test scores (Ladyshevsky, 1996). It is for these reasons that the scores from these students were excluded from this study.

All students were also required to complete a demographic questionnaire asking them about their clinical experiences both within and outside the curriculum. This is displayed in Appendix 8. The investigator was specifically interested in experiences related to shoulder pathology. This information was used to ensure that the IND and RPC groups were similar with respect to clinical experience. Controlling for clinical experience was necessary to ensure that the IND and RPC groups did not differ on this dimension. Even though all students have the same number and type of clinical visits in the undergraduate program, the specific conditions that are encountered

during these visits will vary. Further, many students provide physiotherapy-type services in the community as part of their work with sporting clubs. Hence, they may have gained experience in the management of orthopaedic conditions outside of the program.

Several studies have demonstrated that prior experience with a specific condition, or experience within a certain clinical area may give students an advantage in later evaluations where the same case or content is being examined (Blackwell & Callaway, 1992; Dettmann & Linder, 1988; Haydon et al., 1994; Van Rossum, Briet, Bender, & Meinders, 1990; Woolliscroft, Swanson, & Case, 1992). Figures 6.3 to 6.5 and Tables 6.9 to 6.11 illustrate the equivalency of the IND and RPC groups background experience. As is evident from the three figures and tables, both the IND and RPC groups were relatively equivalent in their exposure to shoulder pathology.

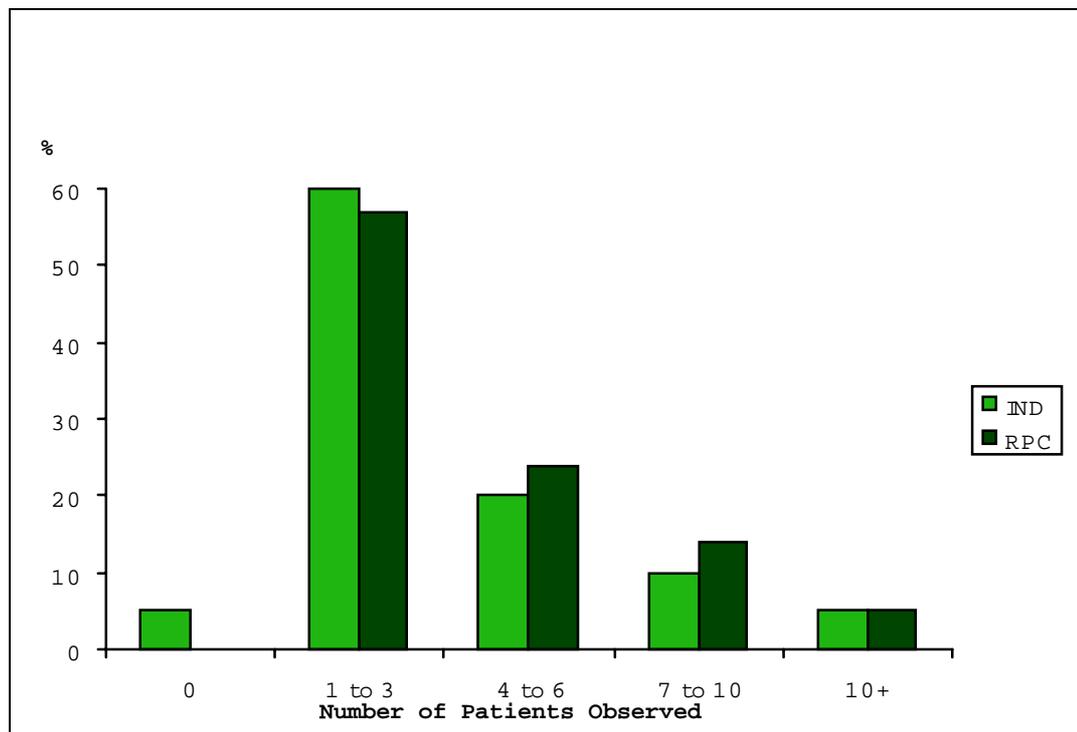


Figure 6.3: Background Shoulder Pathology Experience of IND (n=20) and RPC (n=42) groups (Observation)

Table 6.9: Background Shoulder Pathology Experience of IND and RPC groups (Observation)

Number of Clients Observed	Individual n = 20 Percentage	Reciprocal Peer Coaching n = 42 Percentage
0	5	0
1 to 3	60	57
4 to 6	20	24
7 to 10	10	14
10 or more	5	5

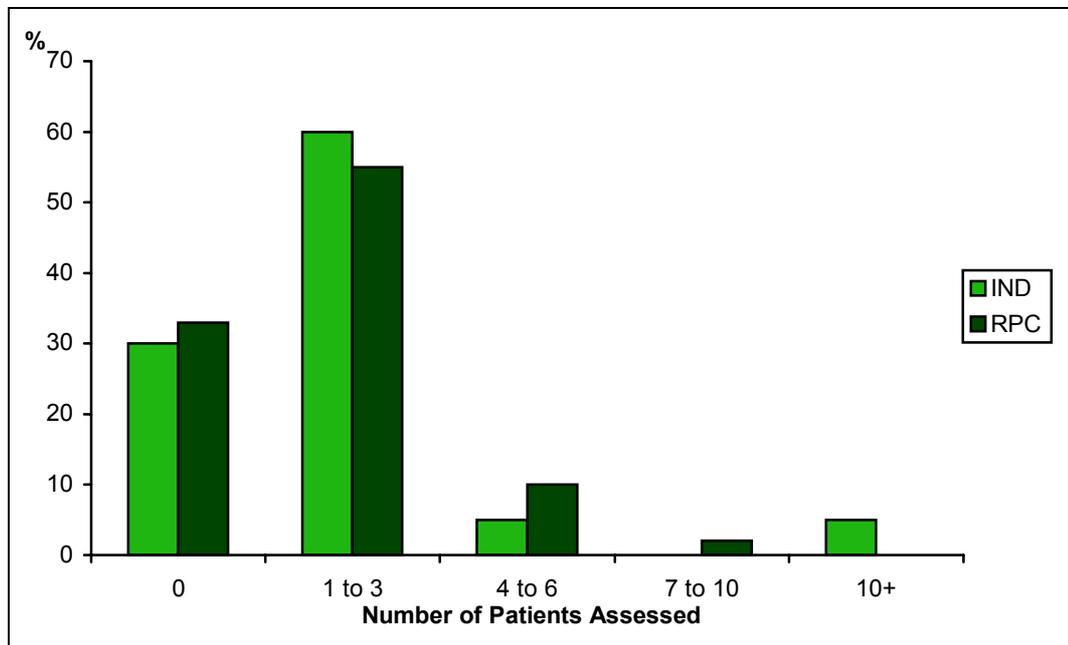


Figure 6.4: Background Shoulder Pathology Experience of IND (n=20) and RPC (n=42) groups (Assessment)

Table 6.10: Background Shoulder Pathology Experience of IND and RPC groups (Assessment)

Number of Clients Assessed	Individual n = 20 Percentage	Reciprocal Peer Coaching n = 42 Percentage
0	30	33
1 to 3	60	55
4 to 6	5	10
7 to 10	0	5
10 or more	5	0

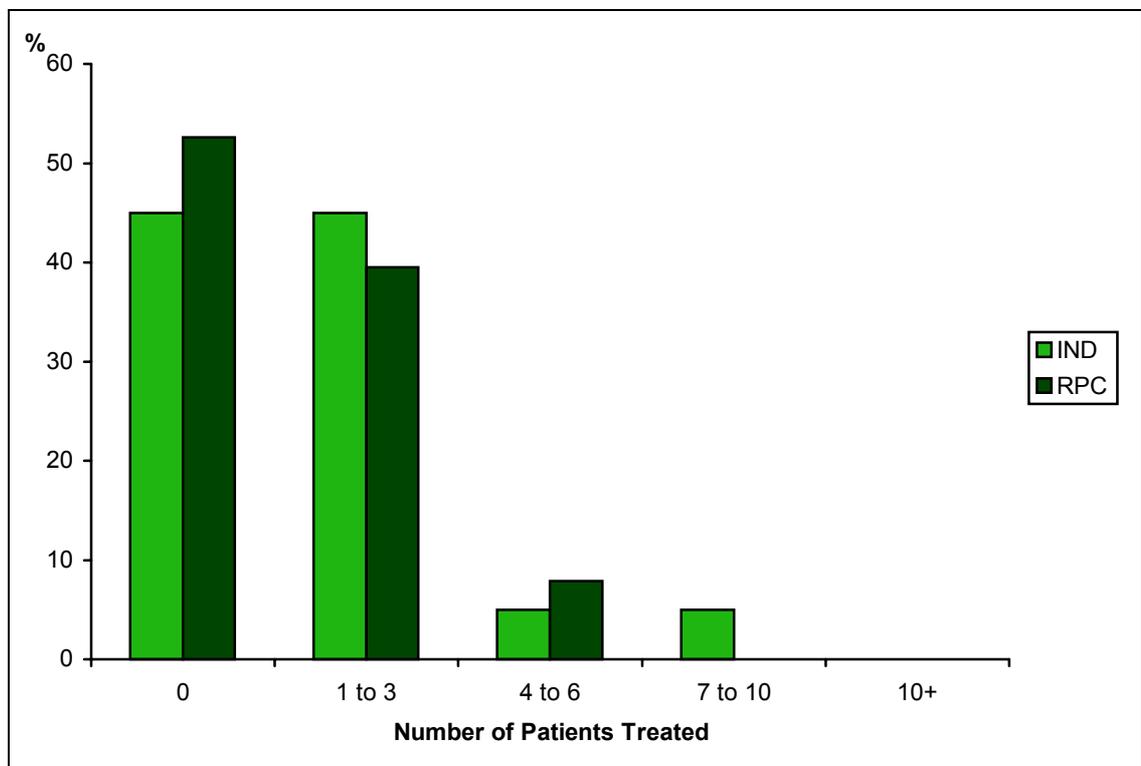


Figure 6.5: Background Shoulder Pathology Experience of IND (n=20) and RPC (n=42) groups (Treatment)

Table 6.11: Background Shoulder Pathology Experience of IND and RPC Groups (Treatment)

Number of Clients Treated	Individual n = 20 Percentage	Reciprocal Peer Coaching n = 42 Percentage
0	45	48
1 to 3	45	43
4 to 6	5	10
7 to 10	5	0
10 or more	0	0

All SP encounters were scheduled outside of class time and took four weeks to complete. The schedule was developed taking into consideration the recommendations of Stillman (1990) and Vu, Steward, and Marcy (1987). They recommend that SPs not portray more than 12 simulations or engage in more than five hours of simulation per day, otherwise, their performance and rating reliability deteriorates.

6.5.3 Grapevine Effect

As was noted earlier, the possibility of a grapevine effect influencing scores (i.e., scores increase on a week by week basis because students talk about the tests) was of concern as the SP encounters took place over a period of four weeks. The issue of sequential testing and examination security was reviewed in detail in Chapter 3 and does not appear to be a major issue in OSCEs. Individual stations within an OSCE, however, have been found to show an increase in scores over time (Colliver et al., 1991). Methods to minimise the possibility of a grapevine effect have been described by Morrison, 1997), Petrusa et al. (1987), Rutala et al. (1991), Williams, Lloyd, & Simonton (1990) and are included in the design of this study. These methods include: having students sign confidentiality statements; reminding them of plagiarism and student code policies, advising them that features of the SP may change over the course of the examination so it is in their best interests to avoid potentially misleading their peers with inaccurate information; keeping all checklist and post encounter questions secure so the students do not know what criteria are being evaluated; and warning students about the possibility of reasoning errors if they manage a case with pre-supposed notions.

Methods to prevent an apparent grapevine effect from occurring also had to be implemented. For example, if most students with high OS scores were scheduled towards the latter half of the study, with lower scoring students going first, then it is possible that the mean scores week by week would increase. This would be due to the method of subject allocation and not a grapevine effect. In light of this, relatively equal numbers of low and high scoring students in both the RPC and IND groups were allocated across the four weeks of testing. For example, eight high achievers were tested each week. There were eight, nine, eight and five low achievers tested across the four weeks respectively. The overall mean academic grades for these student groupings are illustrated in Table 6.12. The mean academic scores of the subject groups are relatively constant across the four weeks of the study.

Table 6.12: Mean Orthopaedic Scores of Subjects Week by Week

Week	n	Mean Score	s.d.
1	16	70.19	9.7
2	17	71.26	9.3
3	16	70.44	7.5
4	13	71.35	7.2

6.5.4 Preparation of Students for Reciprocal Peer Coaching

Students in the physiotherapy program have numerous opportunities to work alongside peers. As noted earlier, many of the clinical visits in second and third year are group based. The tutorial stream in HSBIP also requires students to develop a presentation on a key patient communication strategy as a pair. This presentation is done as a dyad and both students receive the same mark for their contributions. Nonetheless, a specific two hour tutorial on peer coaching was carried out prior to the actual patient encounter for all students enrolled in the HSBIP unit. Students were provided with an article on peer assisted learning (PAL) in clinical education (Lincoln & McAllister, 1993), engaged in a group discussion on the benefits and challenges of PAL, carried out three role plays using a RPC strategy to evaluate a

patient, and discussed communication principles associated with team work and coaching.

6.5.5 The Simulated Patient Encounter

Students in the IND group worked through the SP encounter independently. Students in the RPC group were initially going to be instructed to take a specific role during the SP encounter. This approach followed one taken by Ogden, Barnhart, and Davis (1987) who also used an extended SP as a teaching tool in a PBL course. They had one student take the history while the other student did the physical examination.

During this study's pilot test, however, some of the students in the RPC pairs naturally gravitated towards working collaboratively. History and physical examination developed into shared roles. Other students chose to perform each of these tasks separately, inviting input from their peer coach. The arrangement of the room also influenced how students coached one another. For example, during the pilot test the chairs for one pair of students were set apart with the SP in between. One of the students provided feedback to the investigator that she felt this created an impediment to her being able to coach her peer.

Given these natural variations in how students may work together, the investigator did not impose controls on how the RPC pair should work. The only instruction that was given to the RPC pairs was that which occurred during the peer coaching tutorial. Students were advised they could suggest additional history questions or examination maneuvers during the SP encounter and could call a time out to discuss something related to the management of the case.

The patient encounter took place in an environment that resembled a clinical treatment area. A standard outpatient assessment form was provided to all students in both groups which they could use to guide their evaluation. They were also free to bring in any supplementary notes they had prepared beforehand to guide them through the encounter. This is not unusual as most novices will use preparatory notes to guide them through a new patient encounter. An unobtrusive videocamera recorded the complete encounter. Students had one experience prior to the patient encounter in which their performance was videotaped. This was during their 'patient

communication presentation' to the class as part of the HSBIP unit. The purpose of this videotaping was to give students feedback about their presentation and to get them used to the presence of a videocamera.

Students were instructed that they had up to 60 minutes to complete the patient assessment. They were instructed to leave the assessment area after the history and to plan their physical examination. They had five minutes to complete this task. During this interval, the SP completed the history checklist. The students then returned and completed the physical examination. The time spent on the history and physical examination portions of the assessment were captured for all students.

Twelve observations of the SP's performance were carried out by the investigator who unobtrusively observed the student-SP encounters by positioning himself at the entrance of the treatment cubicle. The investigator also scored a HC, PEC and ACIRS and compared these results with the SP's ratings for ongoing checks of rating reliability. Any discrepancies or concerns were addressed immediately to ensure continuing consistency and accuracy of the simulated patient. Prior to the commencement of the main study, as there was a three month gap between the pilot study and the main study, the SP went through another five hours of rehearsals. The SP's performance accuracy and rating reliability were evaluated once again during this interval.

6.5.6 Post-Encounter Activities

After the patient encounter, the SP completed the PEC and the Arizona Clinical Interview Rating Scale. The students completed the post-encounter anxiety/confidence questionnaire and rated the fidelity of the SP case. This inventory is displayed as Appendix 7. All students were then given up to 15 minutes to think about the case. In the case of the RPC group, they could discuss the case with their peer. All individuals were advised to think about the patient's problems, likely diagnoses and management possibilities as these were going to be the focus of the questions on the post-encounter questionnaire. Several reference texts were available to the students if they needed to follow up a query. Reference texts are often available in the clinical setting for clinicians. As referral to a text is not unusual

practice, these references were made available to the students. After this 15 minute reflection time, or earlier, each student was required to answer the PEQ independently as this made them accountable for understanding the entire case. They were given up to 30 minutes to complete this task. This was the final component of the overall testing process.

6.6 Data Analysis

The quantitative data analyses employed in this study include descriptive statistics and diagrammatical representations of the results, eg. histograms, scatter plots. Where interesting trends emerged, further tests such as effect size estimation were implemented. Where the possibility of an experimental effect appeared to be likely, inferential tests, such as the t-test and ANOVA were applied to the data. The following sections describe these statistical methods in more detail.

The effect size indicator is a simple means analysis and is recommended in research situations where one group of subjects receives a given intervention (in this case the RPC group) and another group does not receive this intervention (Glass, McGaw, & Smith, 1981; Nelson, 1981). The effect-size method evaluates the difference between the means of pairs of treatment conditions and is best divided by the composite group standard deviation thus yielding a standardised mean difference. The advantage of using the composite standard deviation, under the assumption of a common population standard deviation, is that a more precise estimate is possible (Albanese & Mitchell, 1993). Composite standard deviations were employed in this study to calculate effect size.

The effect size indicator (ESI) allows for the examination of the strengths of the relations between the independent and dependent variables and is a useful tool for further analysing the practical significance of results (Glass et al., 1981). This index is calculated using the following formula:

$$ESI = \frac{\text{Mean (Experimental Group)} - \text{Mean (Control Group)}}{\text{composite standard deviation of both groups}}$$

Nelson (1981) and Vockell and Asher (1995) note that the ESI is basically a z-score. A z-score is a standardised score with a mean of 0 and a standard deviation of one. These z-scores range from about -3.00 to +3.00. Hence, the degree to which the average subject in the experimental group is better or worse off than the average control subject can be examined on a percentile basis. For example, “an effect size of .35 means that the experimental group scored .35 standard deviations above the control group in the study under consideration.” p. 357 (Vockell & Asher, 1995). Cohen (1969) classifies effect sizes as small (.20), medium (.50) and large ($\geq .80$).

The F and t tests are inferential tests that measure whether random sampling alone is the reason for group differences (Nelson, 1981). These tests are generally referred to as significance tests. These tests are carried out where practical significance is noted and the investigator is interested in ruling out the possibility that chance alone is the reason for the experimental effects. A test level of $p < .05$ is used in this study to report statistically significant results. All t-tests were calculated using raw data scores.

The one way analysis of variance with post-hoc comparisons was used to see if several groups differed significantly among themselves (Borg & Gall, 1983; Coakes & Steed, 1997). Coakes and Steed (1997) describe this procedure in great detail. The basic procedure is to derive two different estimates of population variance from the data, and then calculate a statistic from the ratio of these two estimates. One of these estimates is ‘between group variance’ and is a measure of the effect of the independent variable combined with error variance. The other estimate is ‘within group variance’ and is a measure of error variance itself. The F ratio is the ratio of between groups variance to within groups variance. A statistically significant F value indicates that the group means are probably not equal. The actual statistically significant difference is detected by conducting post-hoc analysis or t-tests for multiple comparisons (Borg & Gall, 1983). This type of analysis helps to minimise Type I error. Both Tukey’s and Scheffé’s t-tests for multiple comparisons were employed in this study.

The following section outlines the research objectives for this study and the accompanying measurements that were taken to measure these objectives. The specific data analysis procedures follow.

6.6.1 Research Objective 1

To determine differences in undergraduate physiotherapy student performance, from the perspective of a patient encounter, across the IND and RPC learning models.

Research Questions 1: How does history taking skill differ in the IND and RPC groups?

Research Questions 2: How does physical examination skill differ in the IND and RPC groups?

Research Questions 3: How does interviewing skill differ in the IND and RPC groups?

Several scores will be compared for the IND and RPC groups as part of this analysis. These scores are the mean HC, PEC, and ACIRS score, the mean HC + PEC score, and a composite score of all three checklists. These scores are obtained from the SP who rates the students' ability to perform the patient examination. The forms used to calculate these scores can be found in Appendices 1 to 3. In the RPC group, students are evaluated collectively on these checklists and rating scale. Their scores represent the thoroughness of their combined approach. For statistical purposes they are considered to be one 'unit of measurement'.

Research Question 4: Is there a difference in the time it takes to complete the task, thoroughness, and efficiency, across the IND and RPC Groups?

The time taken for students in the IND and RPC groups to complete the history and physical examination was collected as part of the main study. The mean time (MT) to complete the SP encounter was calculated as follows:

$$MT = \frac{\sum_{i=1}^n (tHx + tPEx)}{n}$$

Where:

- tHx = the actual time (in minutes) to complete the history
- tPEx = the actual time (in minutes) to complete the physical examination
- n = number of students

A mean thoroughness score (MTS) (Neufeld, Norman, Feightner, & Barrows, 1981; Rethans et al., 1991) was also calculated for the two groups. This score is a measure of the students' thoroughness in completing the patient evaluation. These scores are obtained from the SP who scores the students' ability to perform the history and physical examination. The MTS was calculated as follows:

$$MTS = \frac{\sum_{i=1}^n (aHC + aPEC)}{tHC + tPEC}$$

Where:

- aHC = the actual history checklist score for each student/dyad
- aPEC = the actual physical examination checklist score for each student/dyad
- n = number of students
- tHC+tPEC = the total possible history and physical examination checklist score

A mean efficiency by time score (METS) (Neufeld et al., 1981; Rethans et al., 1991) was calculated for the two groups. This score looks at the thoroughness of the students' approach (MTS) while taking into consideration the time taken to complete the patient evaluation. The METS is calculated as follows:

$$METS = \frac{\sum_{i=1}^n \{MTS / (tHx+tPEx)\}}{n}$$

Where:

- MTS = the mean thoroughness score for each student/dyad
- n = number of students
- tHx = the actual time (in minutes) to complete the history
- tPEx = the actual time (in minutes) to complete the physical examination

6.6.2 Research Objective 2

To determine differences in clinical reasoning and problem solving across the IND and RPC learning models.

Research Question 5: Are there differences in clinical reasoning across the IND and RPC group from the perspective of diagnostic skill, ability to identify management options for the client, and skill in identifying the key features of the case?

The mean score of the PEQ and its subcomponents was used to evaluate differences in the IND and RPC groups. The PEQ can be analysed as a single measure or as a composite of measures. These measures are described in more detail.

- Mean score on the overall PEQ.
- Mean score for questions 1 and 2 on the PEQ as these questions specifically focus on diagnosis.
- Mean score for question 3 on the PEQ as this question specifically focuses on patient management.
- Mean score for questions 4 and 5 on the PEQ as these questions explore the link between the SP encounter and the student's CR and problem solving skills related to the history.
- Mean score for questions 6 to 8 on the PEQ as these questions explore the link between the SP encounter and the student's CR and problem solving skills related to the physical examination.

The overall mean and sub-component scores of low and high achieving students on the PEQ was evaluated across the IND and RPC groups. The same five analyses described above were conducted.

Another question related to objective two is whether there is a relationship between student performance during the patient encounter (HC and PEC scores) and ability to correctly diagnose and manage a patient. This question is designed to measure whether success in completing the key components of the history and physical examination influence the student's ability to successfully complete the post-encounter questionnaire. This relationship can be explored using scatter plot analyses to see whether there is a direct association between these two variables (Currier, 1990).

The Pearson product moment correlation coefficient method was employed to measure practical significance (Currier, 1990). The Pearson product moment correlation coefficient is an index of the linear relationship between two variables and is expressed as a single numerical value between +1.00 and -1.00 (Currier, 1990). Associations between the HC, PEC and the complete and sub-components of the PEQ were explored.

Research Question 6: Are there differences within the IND and RPC group in terms of their history, physical examination, interviewing and clinical reasoning skills?

Analysis of differences within the RPC and within the IND groups is another question in support of Research Objective 2. There are three dyadic structures within the RPC group: high:high, high:low and low:low. Differences in the various scores by dyadic structure were evaluated using a one way repeated measures analysis of variance with post-hoc comparisons. As there are only two groupings within the IND group: high and low, differences in these scores were evaluated using measures of practical significance and the t-test.

6.6.3 Research Objective 3

To determine differences in the affective domain of clinical competence across the IND and RPC learning models.

Research Question 7: How is anxiety and confidence influenced by the IND and RPC learning models?

Scores for anxiety and confidence were derived from the anxiety and confidence questionnaires displayed in Appendices 6 and 7 respectively. The specific analyses that were carried out are as follows:

- comparison of mean anxiety and mean confidence scores on the pre-test questionnaire within the IND and RPC groups.
- comparison of mean anxiety and mean confidence scores on the post-test questionnaire within the IND and RPC groups.
- comparison of the differences in mean anxiety and mean confidence scores from the pre- and post-test questionnaires across the IND and RPC groups.

6.7 Qualitative Methods

This study also employed qualitative methods to capture and analyse data. These methods are described in the following sections.

6.7.1 Stimulated Recall

Given the theoretical support for verbal protocol analysis outlined in Chapter 4, this qualitative research method was selected for this study. The following sections illustrate the methods that were used in the qualitative component of this study. Qualitative methods were selected in order to capture richer textual information about the CR process and the IND and RPC learning experience.

6.7.2 Stimulated Recall: Sampling Methods

Twelve students participated in this component of the study. The purpose was explained to students at the beginning of the Health and Social Behaviour in Physiotherapy unit and the investigator then selected 12 students randomly from a pool of student volunteers. Informed consent was obtained from each participant.

Six students from the IND group (3 low scorers and 3 high scorers) and six students from the RPC group (3 low scorers, 3 high scorers) were selected by the investigator. For students in the RPC group, two of the low students came from a low/low combination and the other one from a high/low combination. Two of the high students came from a high/high combination and the other one from the high/low combination. Hence, each member of a specific dyadic structure participated in this part of the study.

A sample size of 12 subjects is not considered to be small for a qualitative study using stimulated recall. Patel and Arocha (1995) have stated that qualitative methods are used to describe single episodes in detail rather than trying to determine gross average measures. In light of this, qualitative methods designed to evaluate CR are frequently restricted from using broad samples of cases or clinicians and these constraints are representative of work in the area (Elstein, Shulman, & Sprafka, 1990).

6.7.3 Stimulated Recall: Specific Methods

In order to ensure high quality verbal data, specific instructions were given to all 12 students to maximise level one and level two verbalisations (Ericsson & Simon, 1993). The instructions are described in Appendix 9. Ericsson and Simon (1993) recommend an exercise that subjects should experience prior to participating in a verbal recall session. This exercise was used for all 12 students and is displayed in Appendix 10. The purpose of this exercise was to demonstrate the process of concurrent verbalisation, as most students would not be familiar with this type of talking aloud. The subsequent retrospective recall exercise was to help students structure their reports so they yield the same information as the concurrent verbalisations. Ericsson and Simon (1993) state this practice is necessary, otherwise, subjects may engage in an analysis of why they thought in a certain way.

The literature provides in-depth descriptions of how to use stimulated recall in conjunction with patient simulations as a method for studying the CR process (Barrows, 1987; Elstein et al., 1978; Embrey et al., 1996; Ericsson & Simon, 1993). The stimulated recall session occurred immediately after the completion of the post-encounter testing. The recall sessions were conducted by the investigator. Each student watched their performance on a video playback system and were prompted to talk aloud about what they were thinking at the time. Each student was prompted by the investigator when they were silent for 10 seconds. The prompt that was used was, “what were you thinking?” This constant prompting increases the percentage of information reported by the subject (Bordage & Lemieux, 1986). Otherwise, only the occasional focussed, non-directive question about the students’ thinking and their hypotheses were used by the prompter. This approach has been used when more information is needed to understand the nature of a recall (Barrows, Norman, Neufeld, & Feightner, 1982). The students were also advised to stop the videotape, by pressing the pause button on the remote control, and to verbalise their thoughts. The complete stimulated recall session was captured on audiotape and transcribed by a research assistant with experience in transcribing interview text.

At the end of the recall session for subjects in the RPC group, the dyad was asked to report on their experience in working as a pair. A series of focussed, pre-determined

questions were presented to the dyad. These questions were described earlier and are described in Figure 6.1. The subjects' responses were also recorded on audiotape, transcribed and reviewed for content.

All transcriptions were then reviewed by the investigator while listening to the audiotapes. Any inaccuracies in transcription were corrected. The transcripts were then subjected to a qualitative analysis using methods and principles described in the literature (Barrows et al., 1982; Bordage & Lemieux, 1986; Embrey et al., 1996; Gale & Marsden, 1982; Greenwood & King, 1995; Jensen et al., 1992; Jensen et al., 1990; Miles & Huberman, 1994; Payton, 1985; Thomas-Edding, 1987). These methods were discussed in detail in Chapter 4.

The codes that were used to analyse the textual data were developed in two ways. First, codes that have been used to describe the CR process of health professionals were adopted or modified for use in this analysis (Bordage & Lemieux, 1986; Jensen et al., 1992; Jones, 1995; Rivett & Higgs, 1997). Second, codes to define or describe the text were developed in tandem with the analysis of the data. This approach helps to create codes that are nested in the context of the data and is the 'grounded' approach described by Guba and Lincoln (1981). All codes were reviewed by an occupational therapist with considerable experience in qualitative research to see if they made sense.

Three coding categories were developed to summarise the data from the stimulated recall sessions. The first category codes were specific to CR and are displayed in Table 6.13. These codes were adopted from the literature describing CR in physiotherapy (Jensen et al., 1992; Jones, 1992; Jones, 1995; Rivett & Higgs, 1997). The second category codes relate to generic aspects of the learning experience; for example, learner anxiety, issues related to the simulation, and confidence. These codes are displayed in Table 6.14. The third category codes relate to the RPC experience. These codes are displayed in Table 6.15. The second and third category codes were developed through thematic analysis by repeatedly reading the transcripts and noting emerging themes. These qualitative methods were employed to answer research question eight: what experiences to students report when engaged in an IND and RPC learning model.

Table 6.13: Clinical Reasoning Codes

Code	Description
Source of Symptoms	Structures or processes which are capable of emanating the client's signs, symptoms or dysfunction. eg: "acromion process" "cervical spine" "rotator cuff tendon"
Contributing Factors	Factors which may predispose or may be associated with the development or maintenance of the patient's problem. Can be environmental, behavioural, emotional, physical, or biomechanical. eg: "abduction above shoulder level" "unaccustomed activity" "sustained posture" "sporting activity"
Precautions and Contraindications	Precautions and/or contraindications to physical examination and/or treatment. eg: "irritability and influence on assessment" "stopping short of full movement to minimise pain" "can't put the patient into that position because of pain"
Management of the Client	Indications of aims of management and/or citing treatment options for the client. This may include whether physiotherapy is indicated. eg: "heat or ice might be helpful for him" "will need to treat him right away" "caudal glides would be a good treatment"
Prognosis or Outcome	An estimation of the extent to which the problem appears amenable to physiotherapy. May include the time frame expected for recovery and/or a prediction of the end result of treatment. eg: "should take at least 8 weeks" "not sure how long the problem will take to resolve" "could be a chronic problem"
Mechanisms of Signs & Sympt.	Neural mechanisms which mediate the patient's signs and symptoms: this includes peripheral, central, sympathetic, and affective components. eg: "pain could be referred from the cervical spine" "will need to do neural provocation testing"
Naming or Generating a Hypothesis	Naming a specific diagnostic hypothesis or syndrome eg: "impingement syndrome" "bicipital tendonitis" "rotator cuff tear"
Minimising CR errors	Recognising the need to carry out certain tests or actions or to ask specific questions to minimise the possibility of a reasoning error eg: "will need to do further testing just to be sure" "can't jump to a conclusion yet" "need to ask this just as a precaution"
Psychosocial Inquiry	Noting that their inquiry is directed towards gaining a greater understanding of psychosocial factors which may influence the assessment, management and/or prognosis of the case eg: "want to find out about his social situation" "is there anyone who can help him at home" "he has financial problems and needs to get back to work"

Table 6.14: Learning Experience Codes: Generic

Code	Description
Recognising a knowledge gap	Realising that one is lacking knowledge or uncertain about how to proceed or what to do eg: “do not know what that drug does” “not sure what to do here”
Acknowledging anxiety	Noting the presence of anxiety eg: “I’m feeling anxious” “his pain made me apprehensive” “that made me worried”
Acknowledging a lack of confidence	Noting to oneself that they are lacking in confidence eg: “I do not feel confident” “I just do not feel competent”
Simulation is noted	Noting the situation is a simulation eg: “I knew it was a set up” “that was a peculiar answer” “thinking he was an actor”

Table 6.15: Reciprocal Peer Coaching Codes

Code	Description
Determining Roles	Commenting about or being sensitive to the role of each party in a RPC situation eg: “I felt left out at this point” “We agreed that Mary would do this part” “I wanted to take over”
Affirming a course of action	Noting to oneself that the other party is taking the correct course of action or asking the right question eg: “It was good that she did not do it in 90 degrees of abduction” “I was glad she asked that” “I was thinking along the same track”
Critiquing a course of action	Noting to oneself that the other party is not performing the task correctly eg: “What Mary said was a bit wordy” “Why have they gone to the neck region first?” “I’d do that later with the special tests”
Seeking Clarification	When the observer (coach) notes to themselves that they are lacking information or would like further information eg: “I wanted to add things in” “I wanted to explore the work situation more” “I wanted to go back to the body chart” “I wonder if he has had an xray?”
Intercepting a course of action positively	When one of the pair influences a course of action positively eg: “I was happy that John asked that question” “It was good that Mary cleared that up” “Mike interjected there and got things back on track”
Support	Acknowledging, requesting or taking support from a peer eg: “I was hoping you would take over that” “I wanted him to see that” “better get John to have a go”

The coding of the transcripts were subjected to inter-rater and intra-rater reliability checks. These reliability co-efficients were obtained by comparing the investigator's coding to the coding of a trained research assistant. This research assistant has over 20 years of experience as a physiotherapist. The intra-rater reliability of the investigator was obtained by coding the same piece of textual data two weeks apart. Ten pages of text were used to test coding reliability (Miles & Huberman, 1994). These tests of coding reliability were carried out to ensure that the investigator's interpretations of the textual data were accurate, according to the coding definitions established for this study. Per cent agreement between the investigator's and research assistant's ratings were used to evaluate coding reliability.

The research assistant initially coded ten pages of text after having had an opportunity to review the coding classification scheme for a week. Disagreements or mis-interpretation of coding definitions during this training session were clarified. Following this training session, a first attempt at coding yielded an inter-rater agreement of 49 per cent. Further clarification of coding practice and definitions took place. The second attempt yielded inter-rater agreement in the order of 71 per cent. With further refinement of coding definitions and practice the third coding session yielded an inter-rater agreement of 84 per cent. In each subsequent coding trial, different pages of text were used and included both recall data from IND and RPC group members.

The investigator also re-coded the same transcript data from the third coding session 10 days later to ascertain intra-rater reliability. The purpose of this evaluation was to establish the internal consistency of the investigator as a rater to see if his ratings were stable and consistent over time. Per cent agreement of 87 per cent was obtained on the first trial.

This chapter has outlined the detailed methods used to answer the research aim and objectives for this study. Both quantitative and qualitative methods have been used in this study and are meant to supplement one another. The results of the main study are described in the next two chapters. Chapter 7 describes the quantitative outcomes of this study and Chapter 8 the qualitative outcomes.

Chapter 7: Results

The results of this study are reported in this chapter. No missing values were found. Variable distributions were checked to see whether or not they satisfied the assumptions of the various statistical significance tests undertaken in the univariate analysis. The results of the quality screen can be viewed in Appendix 11. The results are presented under each research objective.

7.1 Research Objective 1:

To determine differences in undergraduate physiotherapy student performance, from the perspective of a patient encounter, across the IND and RPC learning models.

7.2 Research Questions 1 – 3:

Research Questions 1: How does history taking skill differ in the IND and RPC groups?

Research Questions 2: How does physical examination skill differ in the IND and RPC groups?

Research Questions 3: How does interviewing skill differ in the IND and RPC groups?

Table 7.1 outlines the means and standard deviations of the individualistic (IND) and reciprocal peer coaching (RPC) groups' performance on the history, physical examination and patient communication checklists respectively. The RPC group was composed of 21 pairs with a score recorded for each pair. Table 7.1 outlines the range of these scores. In general, students in the RPC group obtained higher scores on all measures. The standard deviations of the RPC group across all categories also demonstrated lower variance in comparison to the IND group. Table 7.1 also reveals a moderate effect size for the history checklist (0.51), in favour of the RPC group. The other categories demonstrate moderate/strong effect sizes (0.67 to 0.82), in favour of the RPC group. This tendency for higher scores in the RPC group was statistically significant across all categories with the exception of the history checklist: physical examination ($t = 2.62$; $df = 39$; $p < .05$); ACIRS ($t = 2.23$; $df = 39$;

$p < .05$); History and Physical Examination ($t = 2.75$; $df = 39$; $p < .05$); and Total Score ($t = 2.85$; $df = 39$; $p < .05$).

Table 7.1: Independent Samples t-tests on History, Physical Examination and Communication Scores between the IND and RPC Groups

Task	Max Score	IND (n=20)		RPC (n=21)		Effect Size	t value
		Mean	s.d.	Mean	s.d.		
History	48	32.05	5.49	34.48	3.83	0.51	- 1.65
Phys. Exam.	40	23.15	6.72	28.14	5.43	0.76	- 2.62*
ACIRS	24	19.45	2.67	21.05	1.83	0.67	- 2.23*
History & Phys. Exam.	88	55.20	9.98	62.62	7.11	0.80	- 2.75*
Total Score	112	74.65	11.52	83.67	8.62	0.82	- 2.85*

* = $p < .05$

Table 7.2: Range of History, Physical Examination and Communication Raw Scores for the IND and RPC Groups

Task	IND (n=20)	Diff.	RPC (n=21)	Diff.
History	16.0 - 40.0	23.0	28.0 - 41.0	13.0
Phys. Exam.	13.0 - 35.0	22.0	15.0 - 37.0	22.0
ACIRS	14.0 - 23.0	9.0	15.0 - 23.0	8.0
History & Phys. Exam.	32.0 - 72.0	40.0	43.0 - 72.0	29.0
Total Score	47.0 - 94.0	47.0	58.0 - 95.0	37.0

The scores identified in Tables 7.1 and 7.2 are also illustrated through the use of boxplots. Coakes and Steed (1999) describe boxplots as a visual system of summarising information about the distribution of scores. Boxplots illustrate such summary statistics as the median, 25th and 75th percentile, as well as extreme scores in the distribution. The lower boundary of the box is the 25th percentile and the upper boundary the 75th percentile. The median is represented by the horizontal line through the centre of the box. Coakes and Steed (1999) state that in order to determine whether a distribution is normal the median should be positioned in the centre of the box. If the median is closer to the top of the box, the distribution is negatively skewed, and if it is closer to the bottom of the box, it is positively skewed. Coakes and Steed (1999) also note that the smallest and largest observed values within the distribution are represented by the horizontal lines at either end of the box. These lines are often referred to as whiskers. If the distribution has extreme scores –

that is three or more box lengths from the upper or lower edge of the box – these will be represented by an asterisk (Coakes & Steed, 1999). Cases with values 1.5 and 3 box lengths from the upper or lower edge of the box are called outliers and are designated by a circle (Coakes & Steed, 1999). The spread or variability of the scores can be determined from the length of the box (Coakes & Steed, 1999).

Boxplots for the history, physical examination and communication checklists are illustrated in Figures 7.1 - 7.3. Again, the higher median performance and low variance of the scores for the students in the RPC group are evident in comparison to the IND group. Interestingly, the physical examination checklist scores are positively skewed for the students in the IND group and more negatively skewed for the RPC group. Only one outlier for the history checklist appeared in the IND group. One outlier for the physical examination checklist and another for the ACIRS appeared in the RPC group.

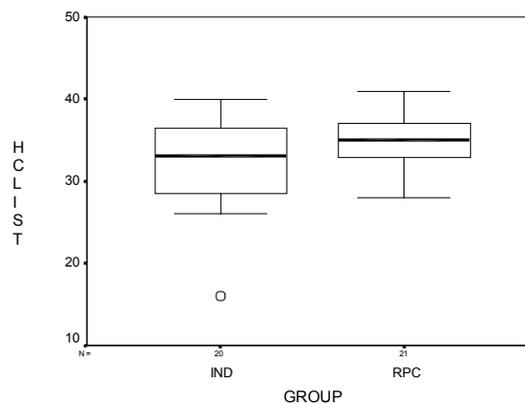


Figure 7.1: Boxplots for History Checklist (HCList): IND and RPC Groups

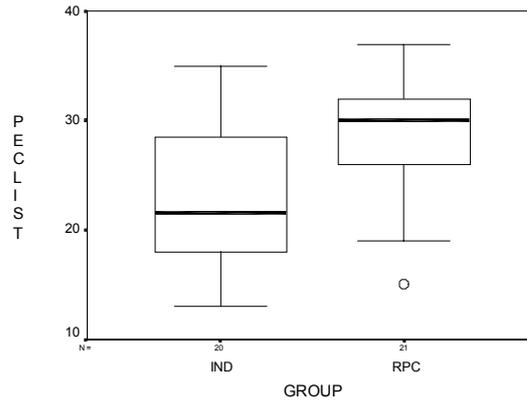


Figure 7.2: Boxplots for Physical Examination Checklist (PECLIST): IND and RPC Groups

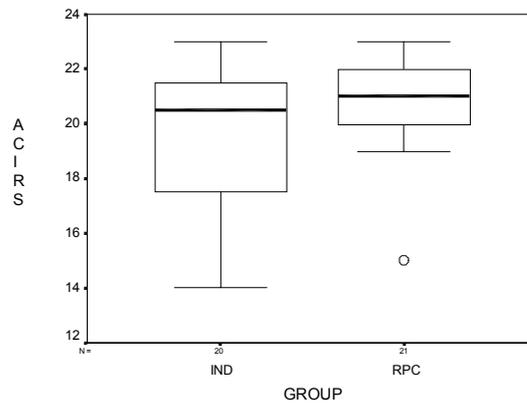


Figure 7.3: Boxplots for the Arizona Clinical Interview Rating Scale (ACIRS): IND and RPC Groups

Figure 7.4 illustrates the mean scores of the IND and RPC groups graphically. They are expressed as a percentage of the total checklist or rating scale score. Subjects in the RPC group outperformed the subjects in the IND group in all categories.

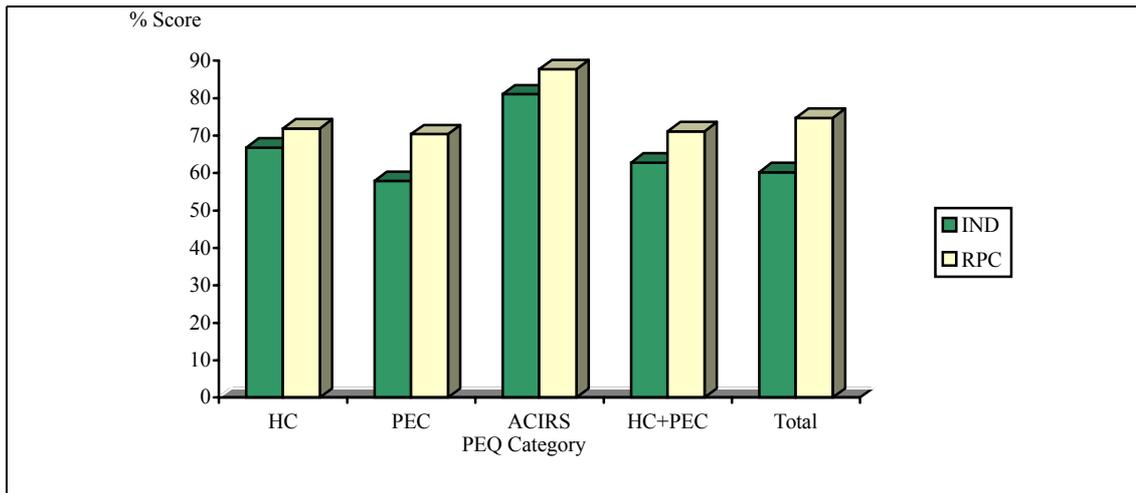


Figure 7.4: History, Physical Examination and Communication Scores for the IND and RPC Groups

7.3 Research Question 4:

Is there a difference in the time it takes to complete the task, thoroughness, and efficiency, across the IND and RPC groups?

7.3.1 Time to Complete the Patient Encounter

Boxplots illustrating the time taken to complete the simulated (SP) encounter are depicted in Figures 7.5 and 7.6. The boxplots illustrate that students in the RPC group required more time to complete the history and physical examination. The variation in time across the two groups is relatively similar. Figure 7.7 illustrates the mean times to complete the patient encounter for the IND and RPC groups graphically.

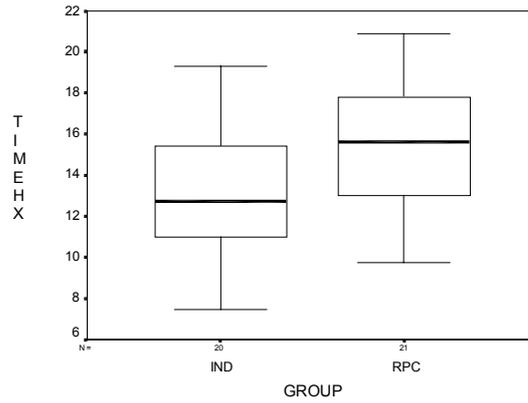


Figure 7.5: Boxplots for Time (in minutes) to Complete the History (TimeHx) for the IND and RPC Groups

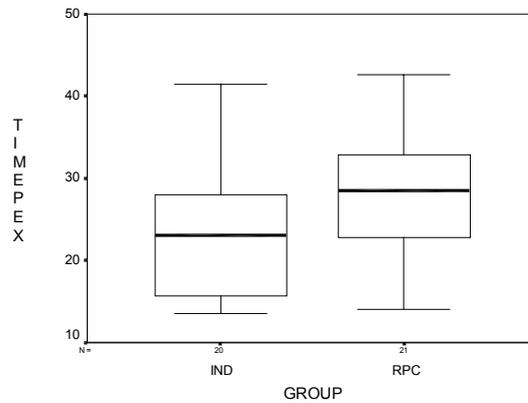


Figure 7.6: Boxplots for Time (minutes) to Complete the Physical Examination (TimePEX) for the IND and RPC Groups

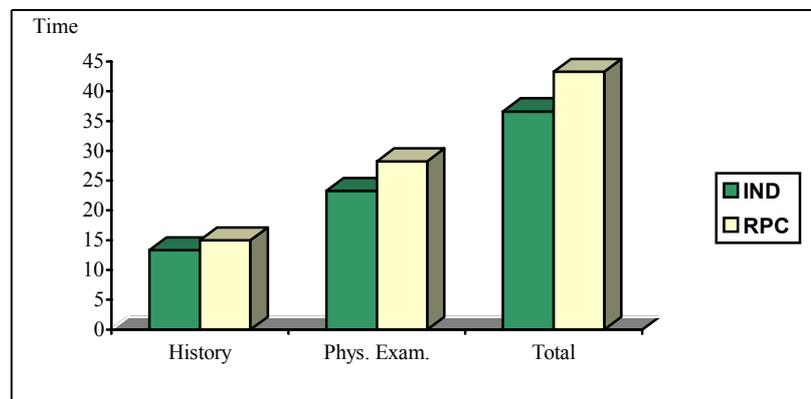


Figure 7.7: Mean Time (in minutes) to Complete the SP Encounter: IND and RPC Groups

Tables 7.3 and 7.4 outline the mean times and range of times to complete the patient encounter for both groups. The RPC group took longer, on average, to complete the history and physical examination in comparison to the IND group. The effect sizes for these differences indicate that this is a moderate effect. Two of these moderate differences were statistically significant with the RPC taking a longer time to complete the physical examination ($t = -1.99$; $df = 39$; $p < .05$) and the total encounter ($t = -2.19$, $df = 39$; $p < 0.05$) in comparison to the IND group.

Table 7.3: Independent Samples t-test for Time to Complete the Patient Encounter: IND and RPC Groups

Task	IND (n=20)		RPC (n=21)		Effect Size	t value
	Mean	s.d.	Mean	s.d.		
History	13.34	3.23	15.03	3.09	0.52	- 1.68
Physical Exam.	23.29	8.41	28.25	7.60	0.60	- 1.99*
Total Encounter	36.63	9.86	43.28	9.54	0.66	- 2.19*

* $p < .05$

Table 7.4: Range of Times to Complete the Patient Encounter: IND and RPC Groups

Task	IND (n=20)	Diff.	RPC (n=21)	Diff.
History	7.47 - 19.33	11.86	9.75 - 20.87	11.12
Physical Examination	13.50 - 41.48	27.98	14.10 - 42.68	28.58
Total Patient Encounter	23.30 - 60.00	36.70	25.57 - 60.00	34.43

7.3.2 Thoroughness of the IND and RPC Groups During the SP Encounter

Boxplots illustrating the thoroughness of the two groups are depicted in Figures 7.8 and 7.9. Thoroughness is measured by dividing actual checklist scores by the total

checklist score. The RPC group was more thorough in completing the history and physical examination. There was also less variation in the thoroughness scores of the RPC group. The scores for physical examination were more positively skewed in the IND group and negatively skewed in the RPC group. These mean differences in thoroughness are illustrated in Figure 7.10.

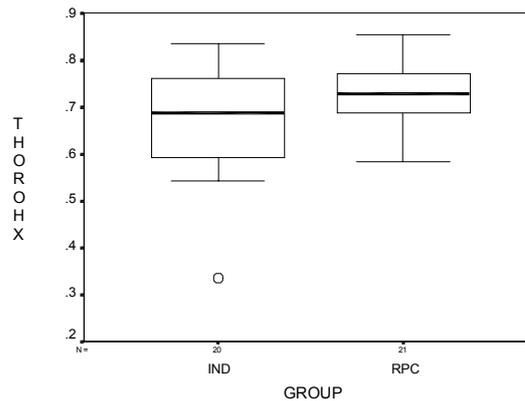


Figure 7.8: Boxplots for Thoroughness of Patient History (ThoroHx): IND and RPC Groups

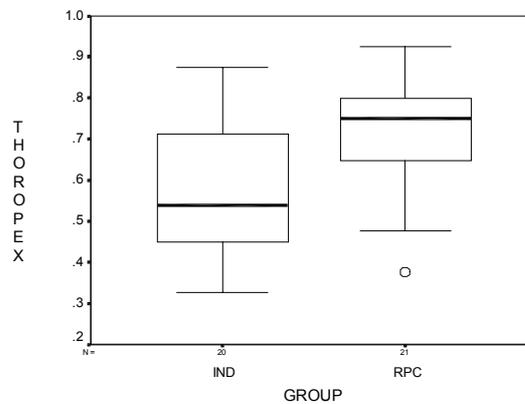


Figure 7.9: Boxplots for Thoroughness of Physical Examination (ThoroPex): IND and RPC Groups

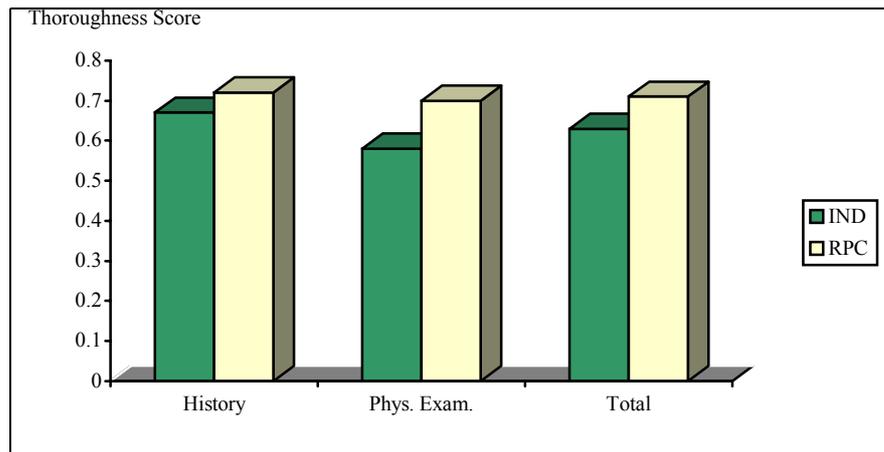


Figure 7.10: Mean Thoroughness of History and Physical Examination: IND and RPC Groups

Tables 7.5 and 7.6 outline differences in the mean thoroughness and range of thoroughness among the IND and RPC groups respectively. The RPC group was more thorough in completing the history, the physical examination and the overall assessment. The effect sizes suggest that the RPC group were moderately more thorough than the IND group on history taking (0.50) and substantially more thorough than the IND group on the physical examination (0.74) and the overall encounter (0.76). The differences in mean thoroughness across the two groups were statistically significant with the RPC group being more thorough on physical examination ($t=2.63$; $df=39$; $p<.05$) and overall assessment ($t=2.75$; $df=39$; $p<.05$).

Table 7.5: Independent Samples t-test for Mean Thoroughness of History and Physical Examination: IND and RPC Groups

Task	IND (n=20)		RPC (n=21)		Effect Size	t value
	Mean	s.d.	Mean	s.d.		
History	0.67	0.11	0.72	0.08	0.50	- 1.65
Physical Examination	0.58	0.17	0.70	0.14	0.74	- 2.62*
Total Patient Encounter	0.63	0.11	0.71	0.08	0.76	- 2.75*

* $p<.05$

Table 7.6: Range of Thoroughness Scores for History and Physical Examination: IND and RPC Groups

Task	IND (n=20)	Diff.	RPC (n=21)	Diff.
History	0.33 - 0.83	0.50	0.58 - 0.85	0.27
Physical Examination	0.33 - 0.88	0.48	0.38 - 0.93	0.55
Total Patient Encounter	0.36 - 0.82	0.46	0.49 - 0.82	0.33

7.3.3 Efficiency of the IND and RPC Groups During the SP Encounter

Tables 7.7 and 7.8 outline the mean efficiency scores and the ranges of these efficiency scores for both groups. The mean efficiency score is derived by taking the thoroughness score and dividing it by the time taken to complete the encounter. As is evident in Table 7.7, there are no appreciable differences across the two groups in terms of their efficiency. The effect sizes are insignificant and there are no appreciable practical or statistically significant differences.

Table 7.7: Independent Samples t-test for Mean Efficiency Scores for History and Physical Examination: IND and RPC Groups

Task	IND (n=20)		RPC (n=21)		Effect Size	t value
	Mean	s.d.	Mean	s.d.		
History	0.052	0.011	0.050	0.012	- 0.18	0.50
Physical Examination	0.026	0.006	0.026	0.004	0.00	0.20
Total Patient Encounter	0.018	0.004	0.017	0.003	- 0.30	0.74

*p<.05

Table 7.8: Range of Efficiency Scores for History and Physical Examination: IND and RPC Groups

Task	IND (n=20)	Diff.	RPC (n=21)	Diff.
History	0.035 - 0.073	0.038	0.030 - 0.075	0.043
Physical Examination	0.018 - 0.038	0.020	0.018 - 0.033	0.015
Total Patient Encounter	0.012 - 0.026	0.014	0.013 - 0.022	0.009

The reason for a lack of difference in efficiency across the two groups can be explained mathematically. Subjects in the IND group were less thorough and took less time to complete the SP encounter in comparison to subjects in the RPC group. Likewise, subjects in the RPC group were more thorough and took longer to complete the SP encounter in comparison to subjects in the IND group. Since the mean efficiency score is the ratio of thoroughness to time taken to complete the encounter, the outcomes for both groups are relatively similar.

7.4 Research Objective 2:

To determine differences in clinical reasoning and problem solving across the IND and RPC learning models.

7.5 Research Question 5:

Are there differences in clinical reasoning across the IND and RPC group from the perspective of diagnostic skill, ability to identify management options for the client, and skill in identifying the key features of the case?

7.5.1 Performance on the Post-Encounter Clinical Reasoning Questionnaire for the IND and RPC Groups

Boxplots illustrating differences in post-encounter clinical reasoning (CR) are depicted in Figures 7.11 -7.15. Performance across the two groups on the diagnosis question was similar (Figure 7.11), with one outlier in the RPC group. Performance on the management question revealed that the RPC group scored higher, with the distribution skewed negatively (Figure 7.12). Median scores for the key feature sections were similar across both groups with interesting variations in skewness and spread (Figures 7.13 – 7.14). For example, students in the IND group had less variation in their answers regarding history key features. There were also three outlier scores, one in the IND group and two in the RPC group. For total score, students in the RPC group had a higher median score with less variation in their performance (Figure 7.15). The differences in mean scores across the two groups are illustrated in Figure 7.16.

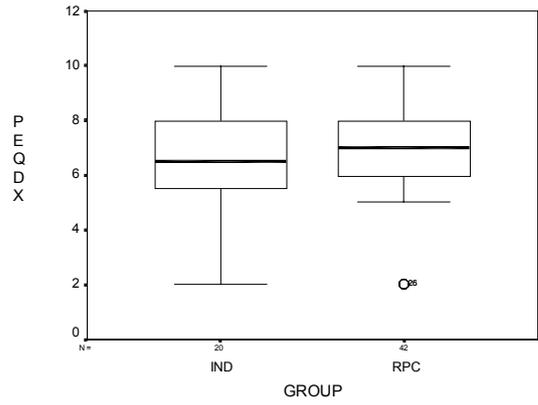


Figure 7.11: Boxplots for PEQ Diagnosis Score (PEQDx): IND and RPC Groups

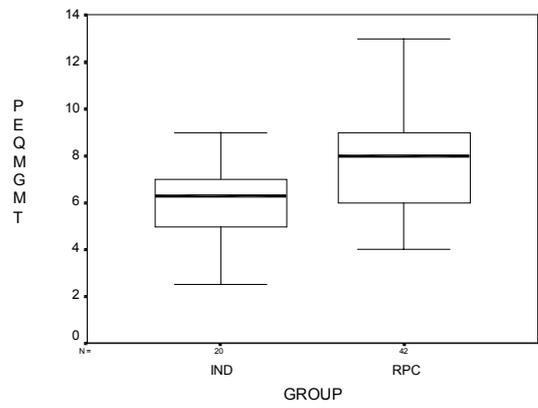


Figure 7.12: Boxplots for PEQ Management Score (PEQMgmt): IND and RPC Groups

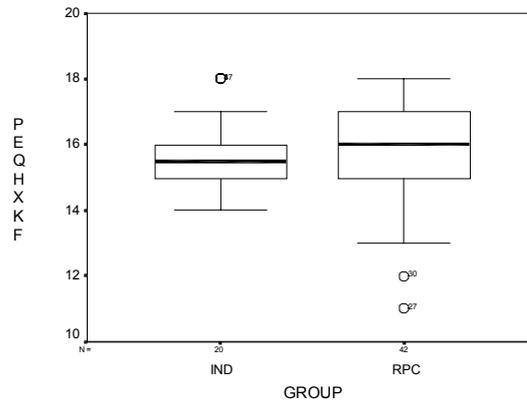


Figure 7.13: Boxplots for PEQ History Key Features Score (PEQHxKF): IND and RPC Groups

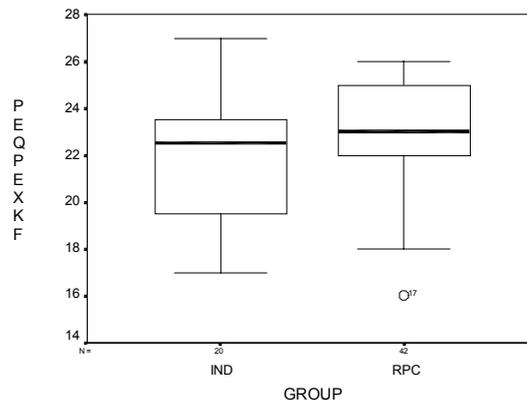


Figure 7.14: Boxplots for PEQ Physical Examination Key Features (PEQPExKF): IND and RPC Groups

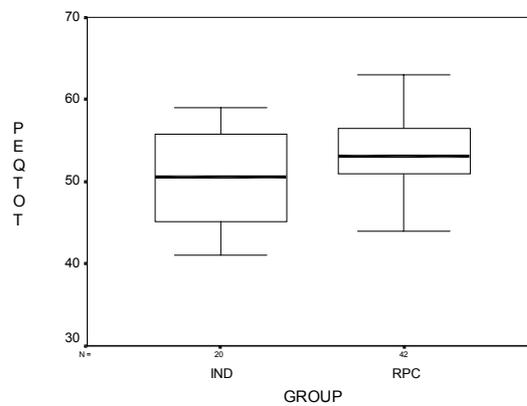


Figure 7.15: Boxplots for PEQ Total Score (PEQTotal): IND and RPC Groups

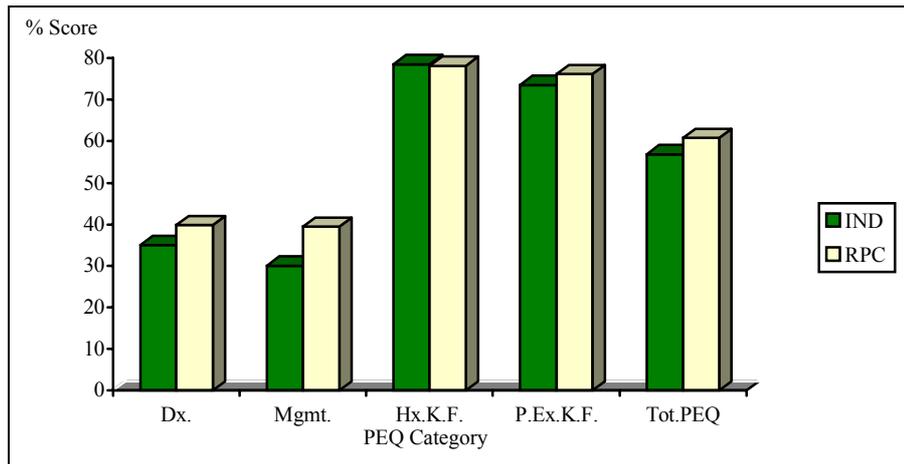


Figure 7.16: Performance Scores on the PEQ for the IND and RPC Groups

Results expressed as a percentage of the total score attainable. Key: Dx=Diagnosis, Mgmt=Management, HxKF= History Key Features, PexKF=Physical Examination Key Features, TotPEQ=Total Post-Encounter Questionnaire

Tables 7.9 and 7.10 outline the mean scores and ranges for the post-encounter CR questionnaire for both the IND and RPC groups respectively. For the most part, the mean scores of the RPC group were higher in comparison to the IND group, with the exception of the history key features mean scores which were relatively equal. Small to moderate positive effect sizes were seen for the RPC group on the diagnosis (0.43) and physical examination (0.32) key features sections of the post-encounter questionnaire. A strong positive effect size was evident for the RPC group on the management section of the post-encounter questionnaire (0.85). A moderate to strong positive effect size for the RPC group occurred for the overall PEQ score (0.69). While the RPC group generally did better on the post-encounter questionnaire, statistically significant differences were only evident for the management section ($t = 3.38$; $df = 60$; $p < .05$) and the total PEQ ($t = 2.66$; $df = 60$; $p < .05$).

Table 7.9: Independent Samples t-test for Outcome Scores for the Post-Encounter Questionnaire: IND and RPC Groups

Items	Max. Score	IND (n=20)		RPC (n=42)		Effect Size	t value
		Mean	s.d.	Mean	s.d.		
Diagnosis	18	6.30	2.45	7.17	1.72	0.43	- 1.61
Management	20	6.00	1.79	7.88	2.16	0.85	- 3.38*
History Key Features	20	15.70	1.26	15.62	1.59	- 0.05	0.20
Phys. Exam. Key Feat.	30	22.05	3.00	22.83	2.15	0.32	- 1.18
Total PEQ Score	88	50.05	5.95	53.50	4.13	0.69	- 2.66*

*p<.05

Table 7.10: Range of Outcome Scores for the Post-Encounter Questionnaire: IND and RPC Groups

Items	IND (n=20)	Diff.	RPC (n=42)	Diff.
Diagnosis	2.0 - 10.0	8.0	2.0 - 10.0	8.0
Management	2.5 - 9.0	6.5	4.0 - 13.0	9.0
History Key Features	14.0 - 18.0	4.0	11.0 - 18.0	7.0
Physical Exam. Key Features	17.0 - 27.0	10.0	16.0 - 26.0	10.0
Total PEQ Scores	41.0 - 59.0	18.0	44.0 - 63.0	19.0

7.5.2 Differences in “Low Student” Performance on the Post-Encounter Clinical Reasoning Questionnaire Across the IND and RPC Groups

Differences in the CR scores for low students in both the IND and RPC group are presented in this section. Boxplots for the post-encounter CR scores are depicted in Figures 7.17 - 7.21. The boxplots for diagnosis reveal that low students in the IND group had much more variation in their scores than their counterparts in the RPC group (Figure 7.17). Their median score was also lower. The median scores for management of the patient were higher for the low students in the RPC group although there was similar variation across the two groups (Figure 7.18). Low students in the RPC group achieved higher median scores for both the history and physical examination key feature sections (Figures 7.19-7.20). Overall, students in the RPC group obtained a higher median score on the overall post-encounter questionnaire with less within group variation (Figure 7.21). These results are illustrated graphically on Figure 7.22 where they are expressed as a percentage.

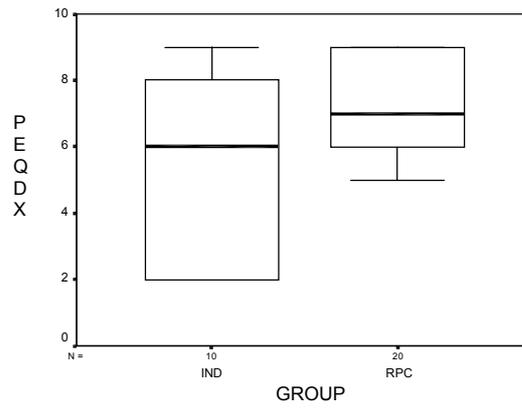


Figure 7.17: Boxplots for Post-Encounter Clinical Reasoning Test: Diagnosis Scores (PEQDx) of Low Students in the IND and RPC Groups

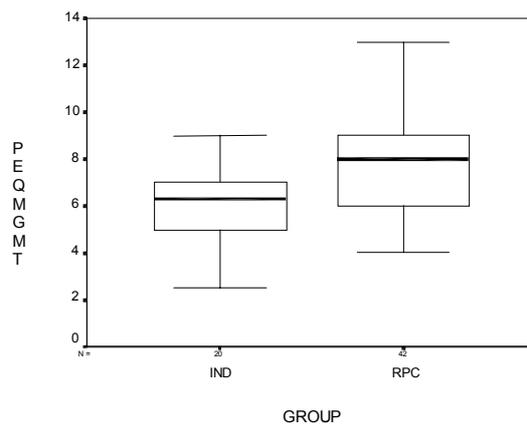


Figure 7.18: Boxplots for Post-Encounter Clinical Reasoning Test: Management Scores (PEQMgmt) of Low Students in the IND and RPC Groups

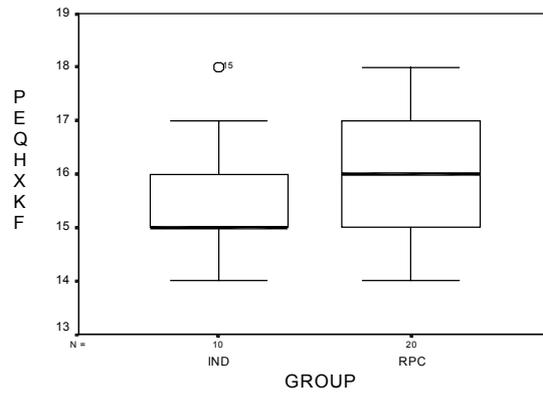


Figure 7.19: Boxplots for Post-Encounter Clinical Reasoning Test: History Key Feature Scores (PEQHxKF) of Low Students in the IND and RPC Groups

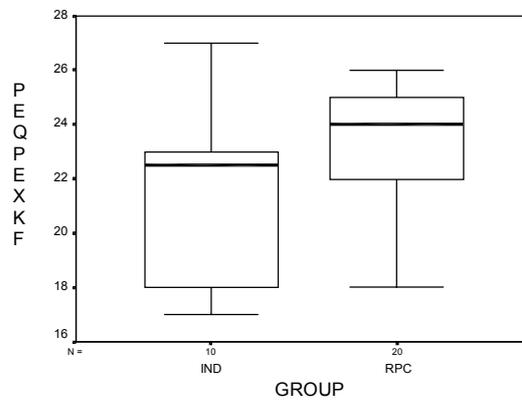


Figure 7.20: Boxplots for Post-Encounter Clinical Reasoning Test: Physical Examination Key Feature Scores (PEQPExKF) of Low Students in the IND and RPC Groups

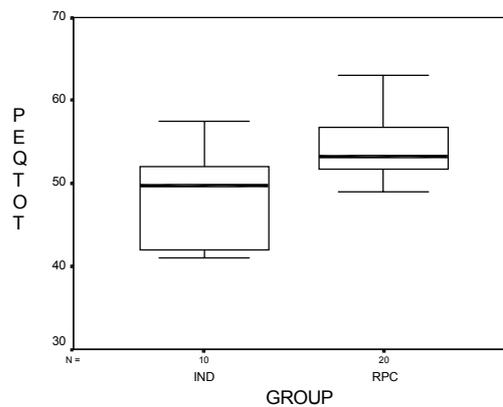


Figure 7.21: Boxplots for Post-Encounter Clinical Reasoning Test: Total Score (PEQTot) of Low Students in the IND and RPC Groups

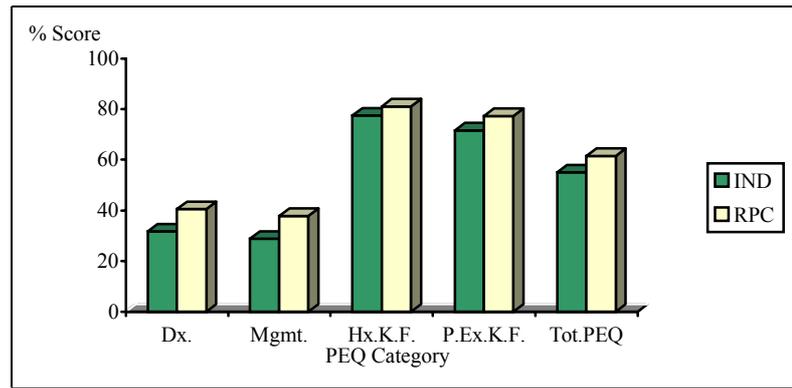


Figure 7.22: Performance Scores on the PEQ for Low Students: IND and RPC Groups

Results expressed as a percentage of the total score attainable. Key: Dx=Diagnosis, Mgmt=Management, HxKF=History Key Features, PexKF=Physical Examination Key Features, TotPEQ=Total Post-Encounter Questionnaire

Tables 7.11 and 7.12 outline the mean scores and range of scores for the low students (by group assignment) on the post-encounter questionnaire. The low students that were assigned to the RPC group scored higher on all aspects of the PEQ in comparison to the low students in the IND group. The extent to which low students in the RPC outperformed low students in the IND group was moderate to strong as demonstrated by the positive effect sizes (0.55 - 1.08). The higher mean scores of the low students in the RPC group were found to be statistically significant for the patient management scores ($t = -2.50$; $df = 28$; $p < .05$) and overall PEQ scores ($t = -3.22$; $df = 28$; $p < .05$).

Table 7.11: Independent Samples t-tests for PEQ Outcome Scores: IND and RPC Groups – Low Students

Category	Max. Score	IND (n=10)		RPC (n=20)		Effect Size	t value
		Mean	s.d.	Mean	s.d.		
Diagnosis	18	5.70	2.87	7.30	1.45	0.75	- 1.66
Management	20	5.75	1.80	7.58	1.92	0.89	- 2.50*
History Key Features	20	15.50	1.27	16.20	1.24	0.55	- 1.45
Physical Exam. Key Features	30	21.50	3.34	23.15	2.13	0.62	- 1.42
Total PEQ Score	88	48.50	5.95	54.23	3.85	1.08	- 3.22*

* $p < .05$

Table 7.12: Range of PEQ Outcome Scores: IND and RPC Groups – Low Students

Items	IND (n=10)	Diff.	RPC (n=20)	Diff.
Diagnosis	2.0 - 9.0	7.0	5.0 - 9.0	4.0
Management	2.5 - 8.0	5.5	5.0 - 12.0	7.0
History Key Features	14.0 - 18.0	4.0	14.0 - 18.0	4.0
Physical Exam. Key Features	17.0 - 27.0	10.0	14.0 - 18.0	4.0
Total PEQ Scores	41.0 - 57.5	16.5	49.0 - 63.0	14.0

7.5.3 Differences in “High Student” Performance on the Post-Encounter Clinical Reasoning Questionnaire Across the IND and RPC Groups

Differences in high student CR across the two groups are presented in this section. Boxplots for the post-encounter CR scores are depicted in Figures 7.23 - 7.27. The boxplots are unremarkable across the two groups except for the post-encounter management question. Here the higher median score for the RPC group is evident. High students in the IND group achieved a higher median history key feature score with less within group variation than the RPC group. These mean scores are depicted graphically in Figure 28.

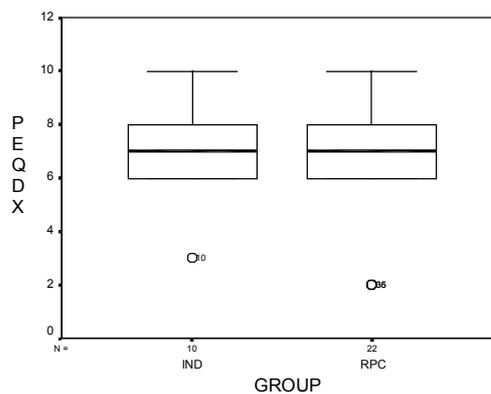


Figure 7.23: Boxplots for Post-Encounter Clinical Reasoning Test: Diagnosis Scores (PEQDx) of High Students in the IND and RPC Groups

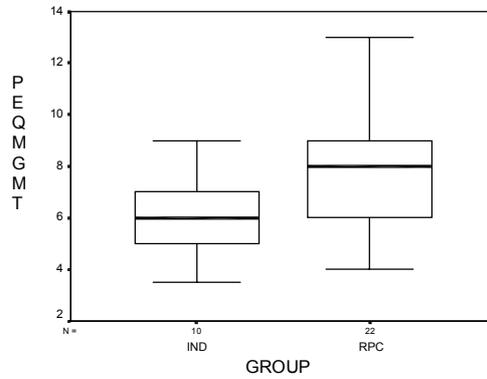


Figure 7.24: Boxplots for Post-Encounter Clinical Reasoning Test: Management Scores (PEQMgmt) of High Students in the IND and RPC Groups

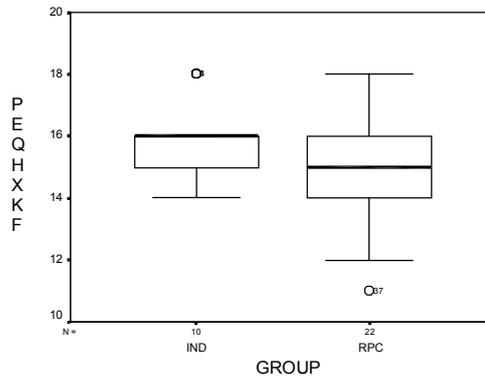


Figure 7.25: Boxplots for Post-Encounter Clinical Reasoning Test: History Key Feature Scores (PEQHxKF) of High Students in the IND and RPC Groups

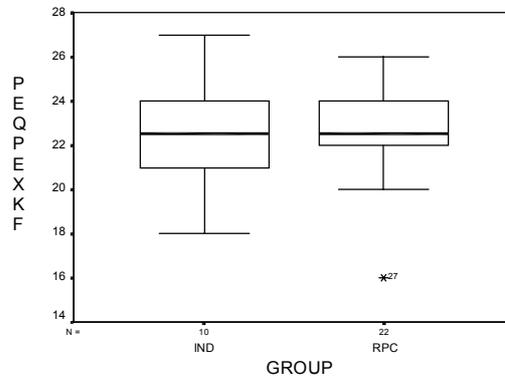


Figure 7.26: Boxplots for Post-Encounter Clinical Reasoning Test: Physical Exam. Key Feature Scores (PEQPEXKF) of High Students in the IND and RPC Groups

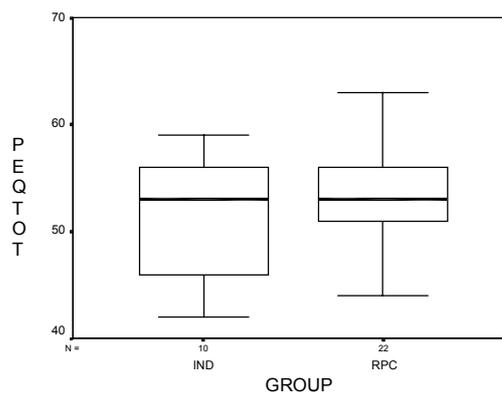


Figure 7.27: Boxplots for Post-Encounter Clinical Reasoning Test: Total Score (PEQTot) of High Students in the IND and RPC Groups

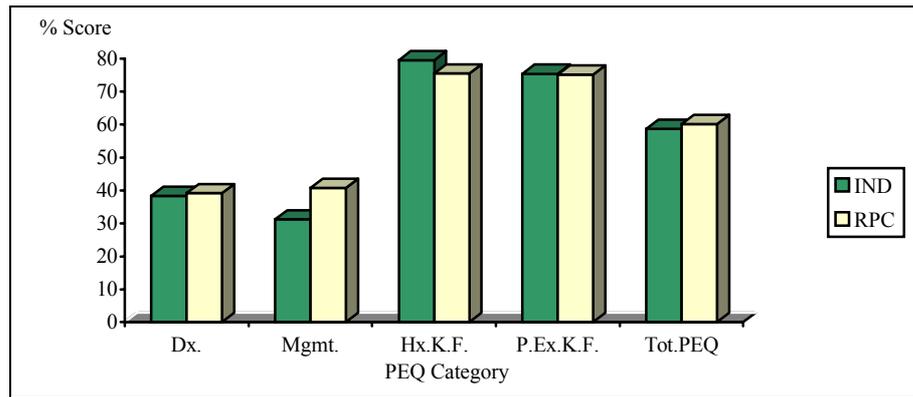


Figure 7.28: Performance Scores on the PEQ for High Students: IND and RPC Groups

Results expressed as a percentage of the total score attainable. Key: Dx=Diagnosis, Mgmt=Management, HxKF=History Key Features, PexKF=Physical Examination Key Features, TotPEQ=Total Post-Encounter Questionnaire

Tables 7.13 and 7.14 outline the mean scores and range of scores for the high students (IND and RPC groups) on the post-encounter questionnaire. With the exception of patient management scores, there were no practical differences across the two groups. Students in the RPC group had higher mean scores on the patient management questions. A strong positive effect size (0.81) in favour of the high RPC students is evident. Further, the higher mean patient management score for the high students in the RPC group was statistically significant ($t = -2.26$; $df = 30$; $p < .05$).

Table 7.13: Independent Samples t-test for PEQ Outcome Scores: IND and RPC Groups – High Students

Category	Max. Score	IND (n=10)		RPC (n=22)		Effect Size	t value
		Mean	s.d.	Mean	s.d.		
Diagnosis	18	6.9	1.9	7.1	2.0	0.08	- 0.20
Management	20	6.3	1.8	8.2	2.4	0.81	- 2.26*
History Key Features	20	15.9	1.3	15.1	1.7	- 0.50	1.33
Phys. Exam. Key Feat.	30	22.6	2.7	22.5	2.2	- 0.04	0.06
Total PEQ Scores	88	51.7	5.8	52.8	4.4	0.25	- 0.65

* $p < .05$

Table 7.14: Range of Outcome Scores for the Post-Encounter Questionnaire: IND and RPC Groups - High Students

Items	IND (n=10)	Diff.	RPC (n=22)	Diff.
Diagnosis	3.0 - 10.0	7.0	2.0 - 10.0	8.0
Management	3.5 - 9.0	5.5	4.0 - 13.0	9.0
History Key Features	14.0 - 18.0	4.0	11.0 - 18.0	7.0
Physical Exam. Key Features	18.0 - 27.0	9.0	16.0 - 26.0	10.0
Total PEQ Scores	42.0 - 59.0	17.0	44.0 - 63.0	19.0

7.6 Research Question 6:

Are there differences within the IND and RPC group in terms of their history, physical examination, interviewing and clinical reasoning skills?

7.6.1 Performance on All Measures Within the RPC Group

Performance measures for high:high, high:low and low:low groups within the RPC sample are presented in this section. Boxplots for the high:high, high:low and low:low groups are illustrated in Figures 7.29 to 7.34. Figures 7.29 to 7.31 illustrate the time taken to complete the history and physical examination and the checklist outcome scores for the three groups. Boxplots for time taken to complete the physical and complete examination clearly illustrate less spread and lower median scores for the low:low group. Figures 7.32 to 7.34 illustrate the outcomes for the PEQ across the three groups.

Table 7.15 outlines the mean scores for all measurement categories associated with the SP encounter. The data are organised by type of group (high:high, high:low, low:low) within the RPC sample. Table 7.16 outlines the mean PEQ scores for the high:high, high:low and low:low pairs.

The eta squared statistic was applied to the data in these analyses to demonstrate practical significance. The eta squared index, also known as the correlation ratio, indicates the proportion of the variance in the dependent variable which is attributable to the independent variable. An eta squared value of 0.10 to 0.15 is generally treated as revealing a strong treatment effect (Kiehl, 1996). Table 7.15 reveals several strong measures of practical significance, suggesting differences in

the independent variable (type of RPC dyad) have an influence on the dependent variables. These practical differences appear in the measures of: time to complete the history (0.11); time to complete the physical examination (.40); time to complete the total SP encounter (.37); history checklist score (.10); physical examination score (.21); and mean efficiency (.37). These differences, particularly those that are statistically significant, are explained further in the next paragraphs.

Students in the high-low group spent significantly less time on the physical examination, [F (2, 18) = 6.02; $p < .05$] and the total encounter [F (2, 18) = 5.38; $p < .05$] in comparison to the high-high and low-low groups. Tukey's HSD test (Kieess, 1996) indicates that the high-low group's mean score on physical examination time (22.30 minutes) was significantly less than the other two groups (32.25 and 31.58). With respect to total examination time, only the difference in mean time between the high:low (36.11) and high:high (48.56) group was statistically significant.

The high-low group also had the highest mean efficiency. This difference in mean efficiency was statistically significant [F (2, 18) = 5.28; $p < .05$]. Tukey's HSD test indicates that the high-low group's mean efficiency (0.019) was significantly higher only in comparison to the high:high group (0.015). The high-low group's higher mean efficiency is the result of this group spending considerably less time on the SP encounter than the other two groups; without a decrease in the high-low group's history and physical examination thoroughness.

Table 7.16 reveals eta squared values for the history key features section (0.27) and composite PEQ score (0.17) which suggest strong practical significance. These differences in the history key feature scores and composite PEQ scores were statistically significant with value of [F (2, 39) = 7.17; $p < .05$] and [F (2, 39) = 3.99; $p < .05$] respectively. When Scheffe's test was applied to the history key feature item, the low:low score (16.83) was significantly higher than the scores obtained by both the high:high (14.79) and high:low (15.44) groups. Scheffe's test is similar to Tukey's HSD test except that it is best applied when data does not present with a normal distribution (Coakes and Steed, 1999). For the composite PEQ score, Tukey's HSD test revealed that only the difference in mean scores between the low:low group (56.13) and the high:high group (52.11) were statistically significant.

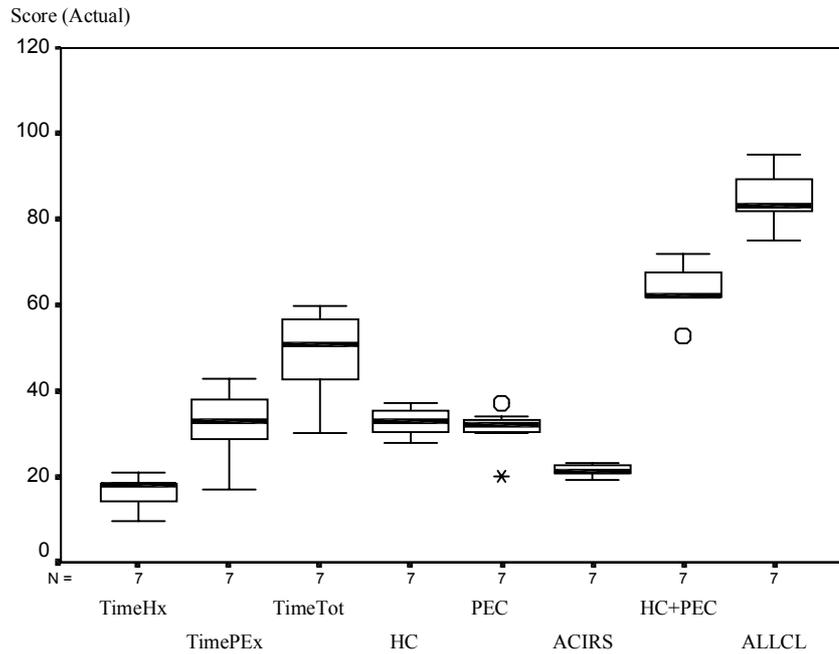


Figure 7.29: Boxplots for the High:High Group - Time and Checklist Measures for the SP encounter

Key: TimeHx=Time History, TimePEX=Time Physical Examination, TimeTot=Time Total, HC=History Checklist, PEC=Physical Examination Checklist, ACIRS=Arizona Clinical Interviewing Rating Scale, HC+PEC=Combined History and Physical Examination Checklist, ALLCL=All Checklists

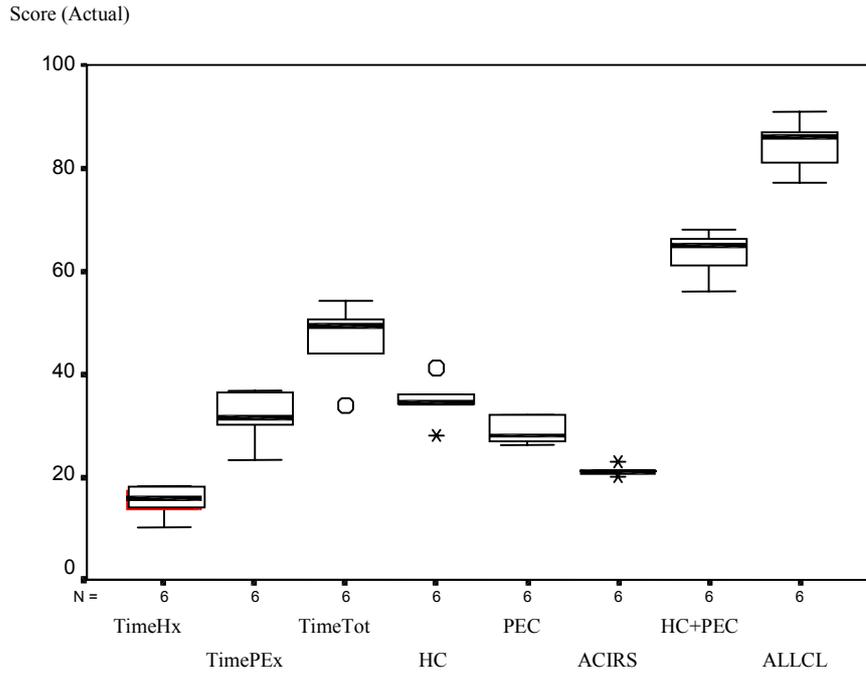


Figure 7.30: Boxplots for High:Low Group - Time and Checklist Measures for the SP Encounter

Key: TimeHx=Time History, TimePEX=Time Physical Examination, TimeTot=Time Total, HC=History Checklist, PEC=Physical Examination Checklist, ACIRS=Arizona Clinical Interviewing Rating Scale, HC+PEC=Combined History and Physical Examination Checklist, ALLCL=All Checklists

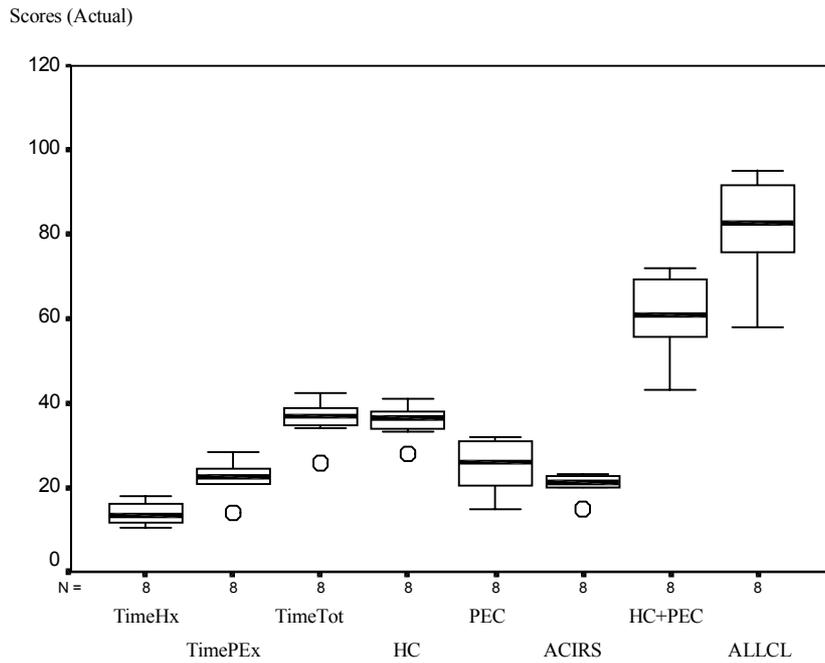


Figure 7.31: Boxplots for Low:Low Group - Time and Checklist Measures for the SP Encounter

Key: TimeHx=Time History, TimePEX=Time Physical Examination, TimeTot=Time Total, HC=History Checklist, PEC=Physical Examination Checklist, ACIRS=Arizona Clinical Interviewing Rating Scale, HC+PEC=Combined History and Physical Examination Checklist, ALLCL=All Checklists

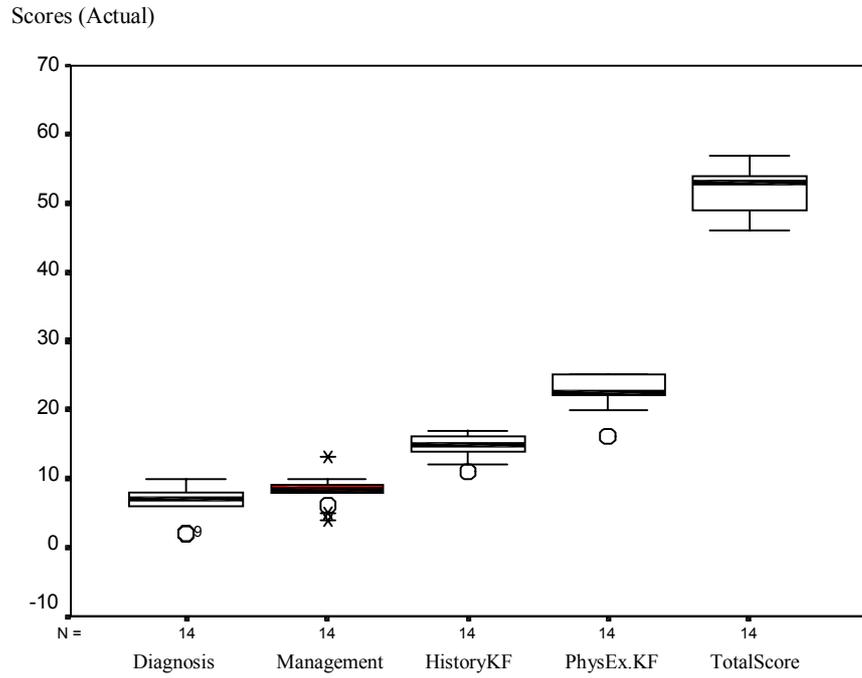


Figure 7.32: Boxplots for High:High Group - PEQ Outcomes

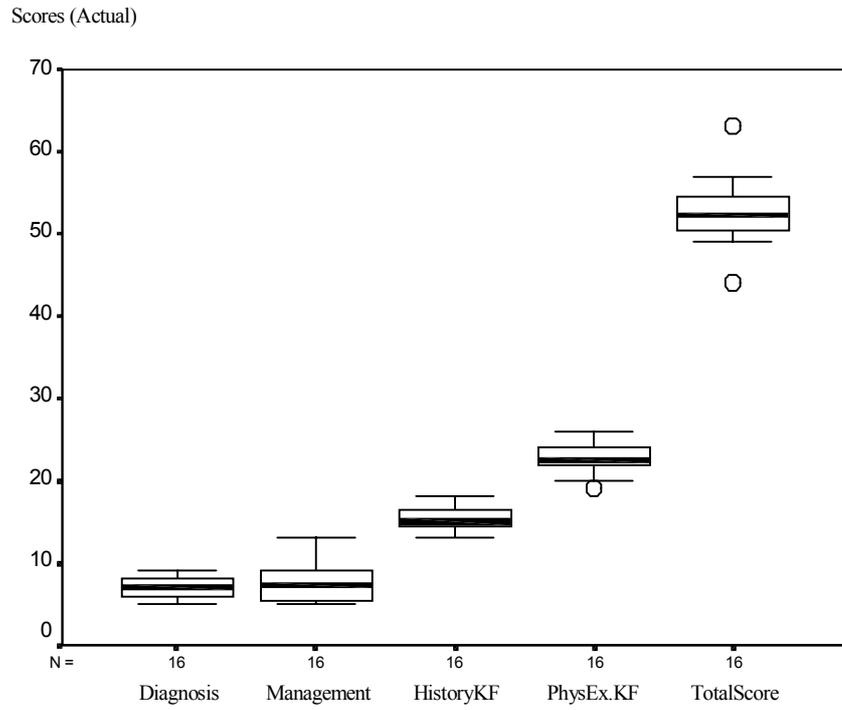


Figure 7.33: Boxplots for High:Low Group - PEQ Outcomes

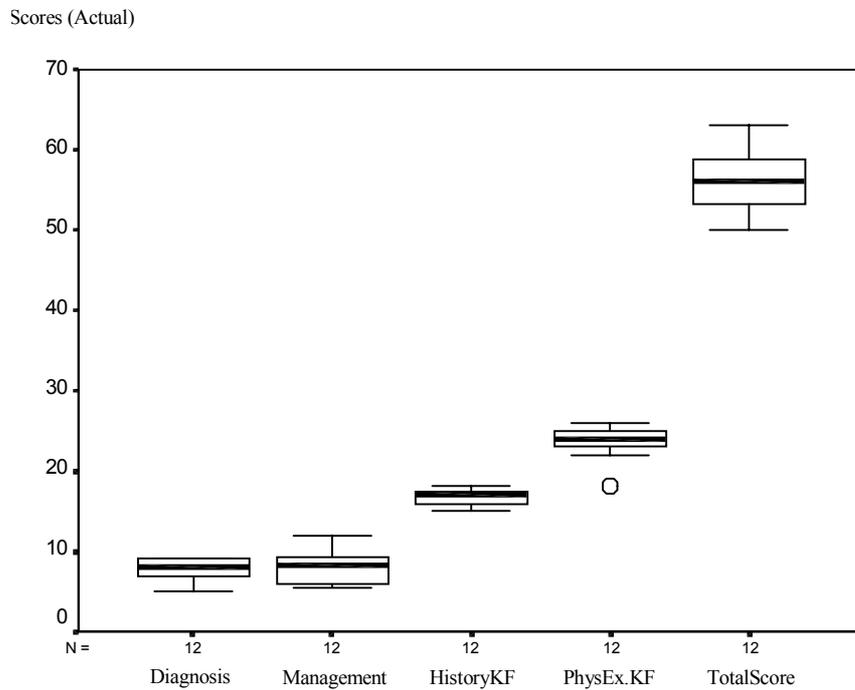


Figure 7.34: Boxplots for Low:Low Group - PEQ Outcomes

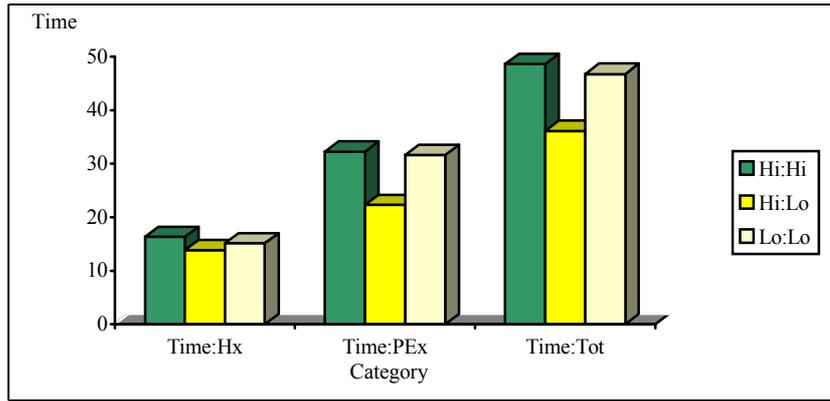


Figure 7.35: Time (in minutes) to Complete the SP Encounter for High:High, High:Low and Low:Low RPC Groups

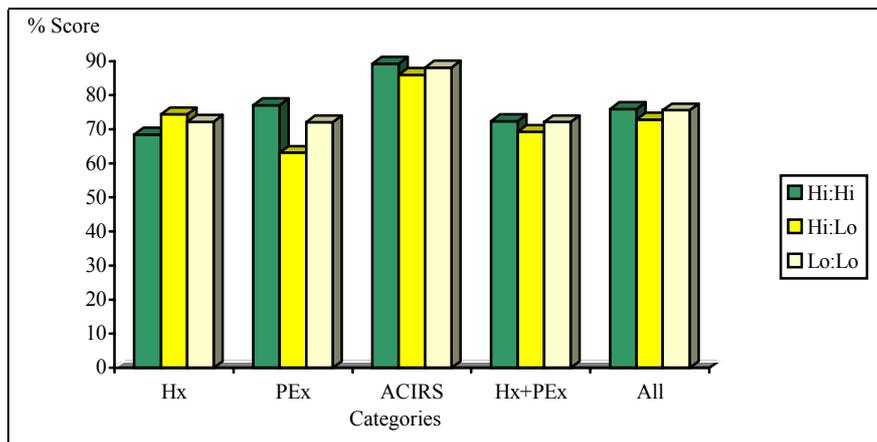


Figure 7.36: Checklist Scores for the SP Encounter for the High:High, High:Low and Low:Low RPC Groups

Results expressed as a percentage of total score attainable.

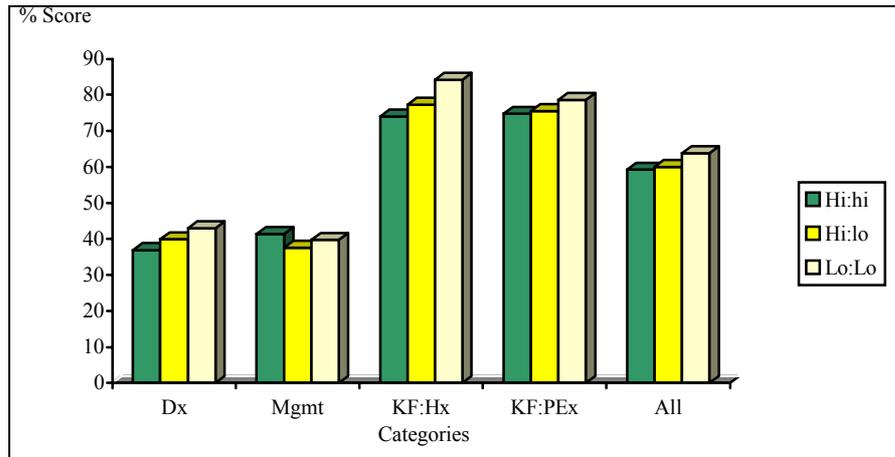


Figure 7.37: Performance on the PEQ for the High:High, High:Low and Low:Low RPC Groups
Results expressed as a percentage of total score attainable.

Table 7.15: One Way ANOVA for Outcomes Scores for the SP Encounter: RPC Group

Category	Max Score	High:High (n=7)		High:Low (n=8)		Low:Low (n=6)		F _{obs}	Eta ²
		Mean	s.d.	Mean	s.d.	Mean	s.d.		
Hx: Time	-	16.32	3.81	13.81	2.61	15.15	3.02	1.18	0.11
P.Ex.Time	-	32.25	8.67	22.30	4.14	31.58	4.88	6.02*	0.40
Total Time	-	48.56	10.82	36.11	4.96	46.73	7.22	5.38*	0.37
History	48	32.86	3.33	35.75	3.92	34.67	4.18	1.09	0.11
Physical Exam	40	30.86	5.31	25.25	6.18	28.83	2.56	2.33	0.21
ACIRS	24	21.43	1.40	20.63	2.62	21.17	0.98	0.35	0.04
History & Phys. Exam.	88	63.71	6.13	61.00	9.68	63.50	4.37	0.31	0.03
Total Encounter	112	85.14	6.79	81.63	12.13	84.67	4.97	0.34	0.04
Mean Thoroughness	-	0.72	0.07	0.69	0.11	0.721	0.05	0.31	0.03
Mean Efficiency	-	0.0154	0.003	0.019	0.002	0.016	0.003	5.28*	0.37

* p<.05

Table 7.16: One Way ANOVA for Outcomes Scores for Post-Encounter Questionnaire: RPC Group

Category	Max Score	High:High (n=14)		High:Low (n=16)		Low:Low (n=12)		F _{obs}	Eta ²
		Mean	s.d.	Mean	s.d.	Mean	s.d.		
PEQ: Diagnosis	18	6.64	2.27	7.19	1.33	7.75	1.36	1.36	0.07
PEQ: Management	20	8.25	2.23	7.50	2.29	7.96	1.98	0.45	0.02
PEQ:History Key Features	20	14.79	1.72	15.44	1.36	16.83	0.94	7.17*	0.27
PEQ: Phys. Exam Key Features	30	22.43	2.44	22.63	1.93	23.58	2.07	1.05	0.05
PEQ:Total	88	52.11	3.60	52.75	4.22	56.13	3.63	3.99*	0.17

*p<.05

7.6.2 Performance on All Measures Within the IND Group

This section provides a description of the differences in the low and high student groupings in the IND sample. Figures 7.38 to 7.40 illustrate differences across these two groups on a variety of measures associated with the SP encounter and post-encounter questionnaire. As is evident from examining these figures, the high students in the IND group outperformed the low students on all of these measures.

Table 7.17 describes the mean outcome scores for all performance measures within the IND group. These scores are grouped into high and low student categories. The effect sizes are moderate to strong for the majority of categories in favour of the high students. Differences in scores that were statistically significant include: physical examination time ($t = 2.87$; $df = 18$; $p < .05$); total encounter time ($t = 3.02$; $df = 18$; $p < .05$); physical examination score ($t = 3.62$; $df = 18$; $p < .05$); history and physical examination score ($t = 2.48$; $df = 18$; $p < .05$); total SP encounter score ($t = 2.47$; $df = 18$; $p < .05$); and thoroughness ($t = 2.48$; $df = 18$; $p < .05$). With respect to the post-encounter questionnaire and its sub-sections, none of the higher scores seen in the high student group were statistically significant.

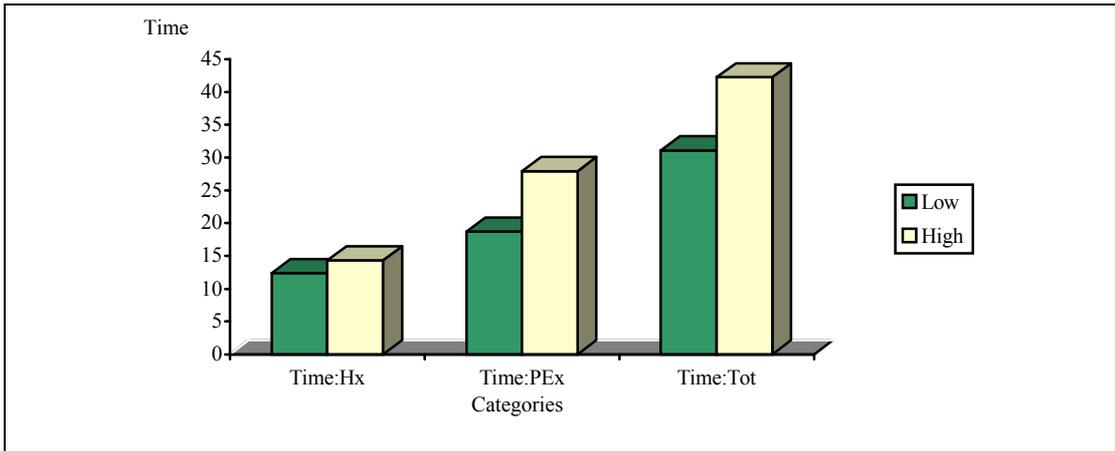


Figure 7.38: Time (in minutes) to Complete the SP Encounter for High and Low Students: IND Group

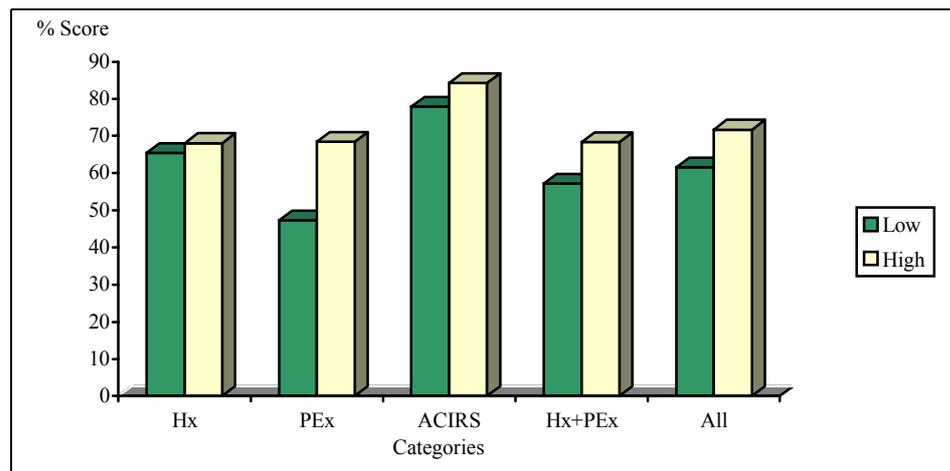


Figure 7.39: Checklist Scores for the SP Encounter for High and Low Students: IND Group

Results expressed as a percentage of total score attainable.

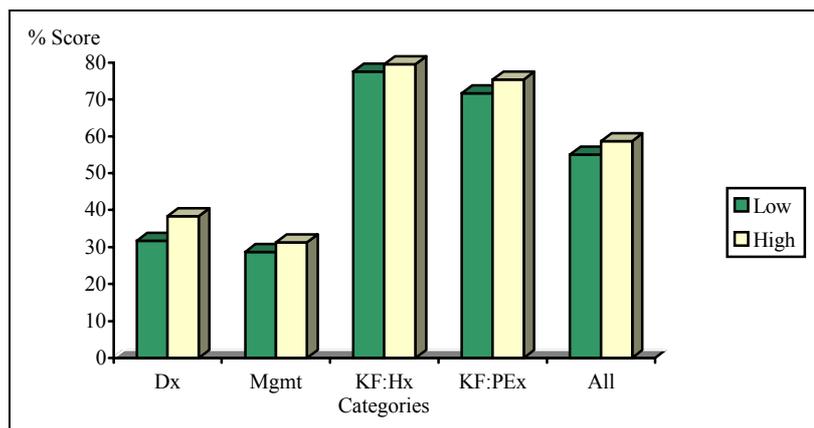


Figure 7.40: Outcome Scores on the PEQ for High and Low Students: IND Group
Results expressed as a percentage of total score attainable.

Table 7.17: Independent Samples t-test for Outcome Scores on All Measures for High and Low Students: IND Group

Category	Max Score	Low Group n = 10		High Group n = 10		Effect Size	t value
		Mean	s.d.	Mean	s.d.		
Time:Hx	-	12.37	2.91	14.32	3.39	0.61	1.39
Time:Phys.Ex.	-	18.69	6.78	27.88	7.51	1.09	2.87*
Time: Total	-	31.06	7.59	42.20	8.86	1.13	3.02*
HC Score	48	31.40	6.50	32.70	4.52	0.24	0.52
PEC Score	40	18.90	4.51	27.40	5.91	1.26	3.62*
ACIRS Score	24	18.70	2.58	20.20	2.66	0.56	1.28
HC+PEC Score	88	50.30	9.49	60.10	8.17	0.98	2.48*
Total Score	112	69.00	10.65	80.30	9.79	0.98	2.47*
Mean Thorough.	-	0.57	0.11	0.68	0.09	0.98	2.48*
Mean Efficiency	-	0.02	0.01	0.02	0.00	-0.26	-0.58
PEQ Diagnosis	18	5.70	2.87	6.90	1.91	0.49	1.10
PEQ Management	20	5.75	1.80	6.25	1.84	0.28	0.61
PEQ Key Feat. Hx	20	15.50	1.27	15.90	1.29	0.32	0.70
PEQ Key Feat. PEx.	30	21.50	3.34	22.60	2.67	0.37	0.81
PEQ Total	88	48.45	5.95	51.65	5.79	0.54	1.22

* p<.05

7.7 The Relationship Between Performance on the Patient Encounter and Post-Encounter Reasoning

The correlations between performance on the patient history and physical examination and the PEQ results are illustrated in Table 7.18. Several statistically significant positive correlations were noted. For example, there were positive correlations between performance on the combined history and physical examination checklist score and performance on the management question of the PEQ ($r = .268$) and overall PEQ ($r = .298$) respectively. Figures 7.41 and 7.42 provide a scatter plot analyses of these results with the accompanying regression line revealing these positive correlations.

Table 7.18: Correlations between the History and Physical Examination Checklists and the Post-Encounter Questionnaire

	HC	PEC	HCPEC	PEQ Diag.	PEQ Mgmt.	PEQHx Key Feat.	PEQPEx Key Feat.	PEQTot Score
HC	1.000	.304		-.018	.128	.116	.214	.189
PEC	.304	1.000		.160	.283	.102	.125	.281
HCPEC			1.000	.105	.268	.133	.200	.298
PEQ Diag.	-.018	.160	.105	1.000	.025	.003	.398	
PEQ Mgmt.	.128	.283	.268	.025	1.000	.028	.205	
PEQHx Key Feat.	.116	.102	.133	.003	.028	1.000	.183	
PEQPEx Key Feat.	.214	.125	.200	.398	.205	.183	1.000	
PEQTot Score	.189	.281	.298					1.000

Critical value for the statistical significance of these correlation coefficients at the .05 level is .250

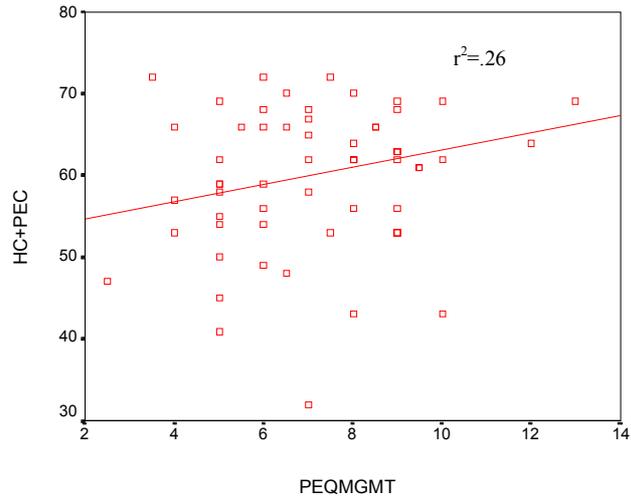


Figure 7.41: Scatterplot of Relationship between History and Physical Examination Checklist Score to Post-Encounter Questionnaire: Management Question Score

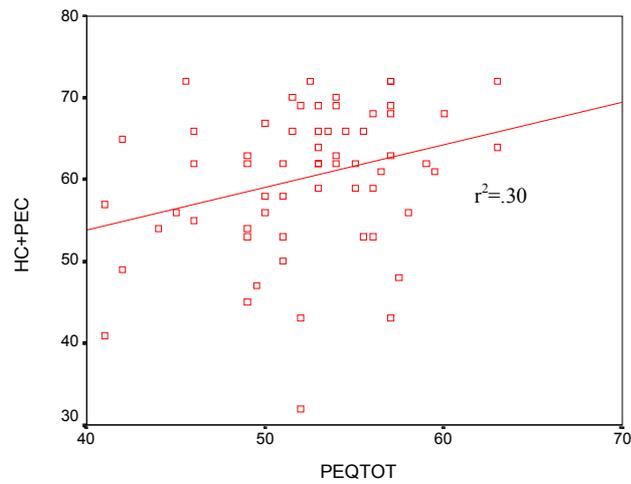


Figure 7.42: Scatterplot of Relationship between History and Physical Examination Checklist Score to Overall Post-Encounter Questionnaire Score

7.8 Research Objective 3

To determine differences in the affective domain of clinical competence across the IND and RPC learning models.

7.9 Research Question 7:

How is anxiety and confidence influenced by the IND and RPC learning models?

7.9.1 Anxiety and Confidence Scores of the Subjects

Tables 7.19 and 7.20 outline the pre-encounter and post-encounter anxiety and confidence measures for the IND and RPC groups respectively. These scores were anchored on a five point Likert scale with low scores indicating that the subjects were very anxious and lacking confidence. High scores indicate students were not anxious and very confident.

Tables 7.19 and 7.20 reveal that subjects in both groups became less anxious as they worked through the patient encounter. This tendency to become less anxious is illustrated graphically in Figure 7.43. In the case of the RPC group, there was a strong effect size (0.81) indicating that students in this group became less anxious. This reduction in mean anxiety level over the course of the encounter was statistically significant for the RPC group ($t = -4.23$; $df = 41$; $p < .05$). While students in the IND group also became less anxious, the effect size was small (0.29) and the degree of change was not statistically significant.

Tables 7.19 and 7.20 also reveal that subjects in both groups became less confident as they worked through the patient encounter. The decrease in confidence in the RPC group, however, was smaller than the decrease in confidence for the IND group. This is depicted graphically in Figure 7.44. Correspondingly, the effect sizes for decreasing confidence in the RPC and IND groups were small and small-to-moderate respectively. Results from the t-test revealed that none of these downward changes in either the RPC or IND groups' confidence levels were statistically significant.

Two-tailed independent samples t-tests did not indicate any statistically significant differences in mean pre-encounter anxiety, post-encounter anxiety, pre-encounter confidence or post-encounter confidence between the IND and RPC group. The

independent t test scores for pre-encounter and post-encounter anxiety across the IND and RPC groups were ($t = 0.58$) and ($t = -1.40$) respectively. The t-scores for pre-encounter and post-encounter confidence across the IND and RPC groups were ($t = -0.48$) and ($t = -1.10$) respectively. None of these scores exceed the critical value of $t_c(.05, 60 \text{ df}) = 2.00$.

Table 7.19: Paired Sample t-tests for Anxiety and Confidence Measures for the IND Group

	n	Pre-Encounter		Post-Encounter		Effect Size	t value
		Mean	s.d.	Mean	s.d.		
Anxiety	20	2.58	0.91	2.83	0.85	0.29	-1.23
Confidence	20	3.18	0.63	2.90	0.72	-0.40	1.40

1 = High Anxiety, 5 = No Anxiety; 1 = No Confidence, 5 = High Confidence

Table 7.20: Paired Sample t-tests for Anxiety and Confidence Measures for the RPC Group

	n	Pre-Encounter		Post-Encounter		Effect Size	t value
		Mean	s.d.	Mean	s.d.		
Anxiety	42	2.45	0.71	3.15	0.87	0.81	-4.23*
Confidence	42	3.25	0.55	3.14	0.84	-0.15	0.86

1 = High Anxiety, 5 = No Anxiety; 1 = No Confidence, 5 = High Confidence
 $p < .05$

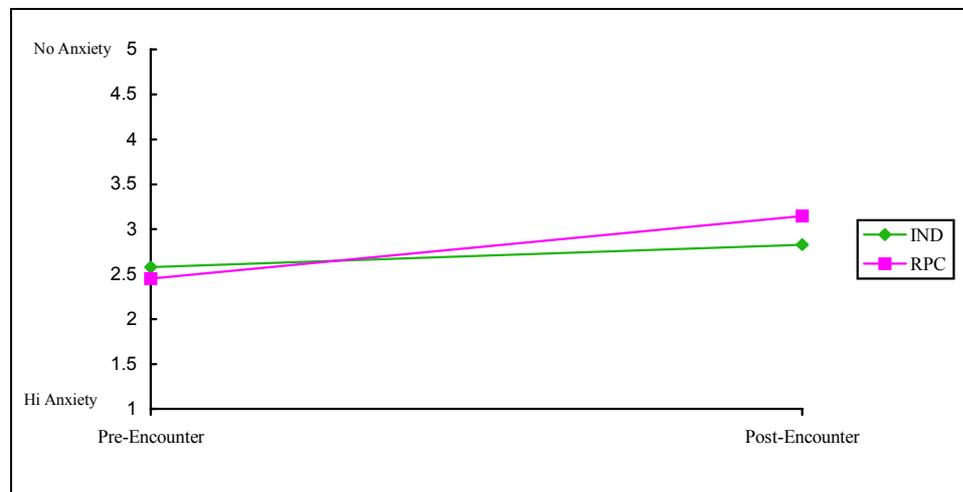


Figure 7.43: Pre- and Post- Anxiety Scores for Both Groups

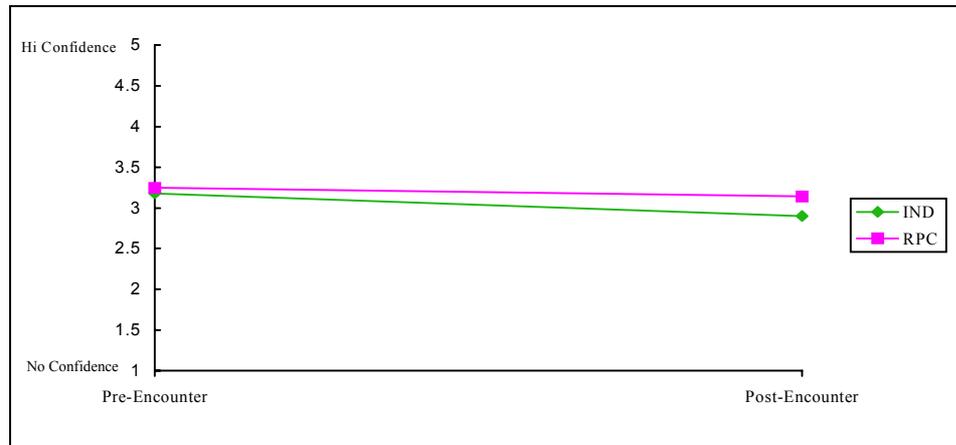


Figure 7.44: Pre- and Post- Confidence Scores for Both Groups

7.10 Features of the Simulated Patient Case

Another minor objective of the study was to determine whether a full scale patient simulation can be used with confidence in studies of clinical reasoning and performance in physiotherapy. This is addressed in the following sub-sections.

7.10.1 Accuracy of the Simulated Patient (SP)

The SP proved to be very accurate in the study. Of the twelve SP encounters observed by the investigator, history findings were presented accurately 98.7 per cent of the time and physical examination findings were presented accurately 93.9 per cent of the time.

7.10.2 Inter-rater Reliability of the Simulated Patient

The SP demonstrated good reliability in rating the subjects on the history and physical examination checklist and the Arizona Clinical Interviewing Rating Scale. The SP's ratings were compared to the chief investigator's ratings. Overall, the SP's inter-rater reliability, expressed as percent agreement, was 91.2 per cent. Expressed as a Kappa statistic, the SP's overall rating reliability was $K=(0.84)$. Table 7.21 summarises the percent agreement and Kappa statistics.

Table 7.21: Inter-rater Reliability of the Simulated Patient

Category	Per cent Agreement	Kappa Co-efficient
History	92.7	.88
Physical Examination	92.1	.86
History and Phys. Exam.	92.4	.86
ACIRS	86.8	.69
All Categories	91.2	.84

7.10.3 Grapevine Effects

Given that this study evaluated students' performance over the course of four weeks, it is quite possible that subjects may have disclosed features of the SP encounter to other students. This disclosure was minimised, in part, by implementing a variety of controls (see methods and review of literature). Nonetheless, student disclosure of the case could have a deleterious effect on scores obtained later in the study. This is known as a grapevine effect (Williams 1987, 1990). One way of measuring a grapevine effect is to compare the results of subjects at pre-determined intervals to see if there are any statistically significant linear trends. In the case of this study, subjects' scores were grouped on a week by week basis and compared. This created four groups for comparison. The weekly groups were organised in such a way that there were equivalent numbers of IND and RPC groups and equivalent numbers of high and low achievers. Table 7.22 describes the composition of these groups.

Table 7.22: Group Composition Structure Across the 4 Weeks

Week	Individuals (n)	RPC Dyads (n)	High Achievers (n)	Low Achievers (n)
------	--------------------	------------------	-----------------------	----------------------

1	6	5	8	8
2	5	5	8	9
3	6	5	8	8
4	3	5	8	5

Tables 7.23 and 7.24 outline the mean scores for the various measurement instruments (history checklist, physical examination checklist, combined history and physical examination checklist and post-encounter questionnaire) on a week by week basis. In general, the mean scores in all categories tended to decrease or stay constant over the course of the study. This is illustrated in Figure 7.45. In spite of this tendency for scores to decrease somewhat, there were no statistically significant differences across any of the mean scores throughout the duration of the study, except in the case of mean history checklist scores [$F(3, 37) = 3.19; p < .05$]. Tukey's HSD test indicates that there was a statistically significant decrease in the mean history checklist scores between the first and third weeks (week 1 = 36.00; week 3 = 30.18). Overall, if collusion between subjects was taking place over the course of the study, it did not appear to have widespread effects on mean performance scores.

Table 7.23: History and Physical Examination Checklist Scores Across the Duration of the Study

Category	Max Score	Week 1 (n=11)	Week 2 (n=11)	Week 3 (n=11)	Week 4 (n=8)	F _{obs}
Mean HC	48	36.00	33.82	30.18	33.13	3.19*
Mean PEC	40	26.09	27.45	25.18	23.50	0.59
Mean HC+PEC	88	62.09	61.27	55.36	56.63	1.40

p<.05

Table 7.24: Post-Encounter Questionnaire Score Across the Duration of the Study

Category	Max Score	Week 1	Week 2	Week 3	Week 4	Fobs

		(n=16)	(n=17)	(n=16)	(n=13)	
Mean PEQ	88	53.09	51.65	53.63	50.92	0.916

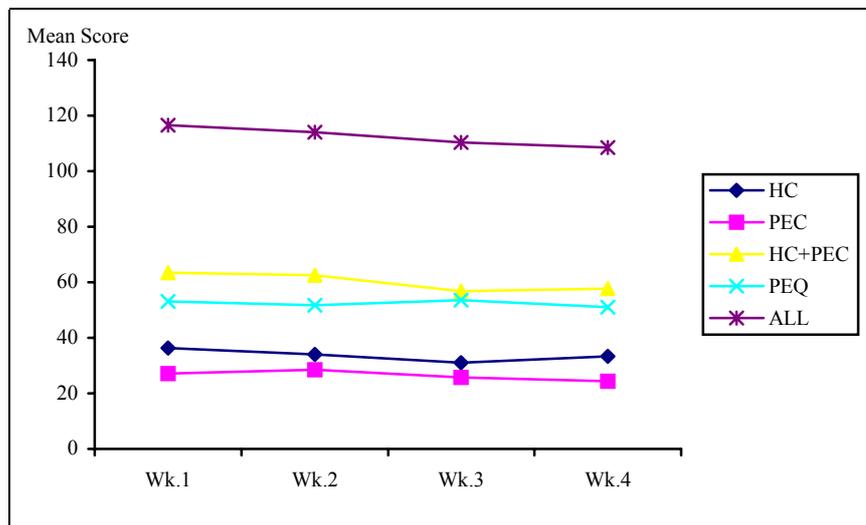


Figure 7.45: Linear Representation of Weekly Scores

7.11 Research Question 8:

What experiences do students report when engaged in an IND and RPC learning model?

The data that are presented in this section are descriptive in nature and meant to supplement the quantitative findings presented earlier in this Chapter. In light of the small sample size in the qualitative part of this study, however, one must exercise caution in making any generalisations from the findings that emerge from this analysis. The phenomena that are presented, however, do provide further insights into the CR processes of the novice physiotherapist and contribute further to the research on peer assisted learning (PAL) in the clinical education setting.

The data reduction strategies which were described in Chapter 4 have been utilised to analyse the qualitative data. As was noted earlier, some of these codes have been taken from the literature on CR in physiotherapy. Other codes were developed using the grounded approach and evolved, were restructured, linked, or interrogated in

different ways to yield the final coding definitions for this study. Tactics for extracting meaning from this data have included the presentation of themes or patterns, clustering, counting and making contrasts and comparisons. Tactics for confirming these findings have been employed by looking for representativeness across subjects, looking for replications across subjects and including outliers or phenomena that do not necessary follow the main thematic representations. The qualitative data are presented in three sections: clinical reasoning, the learning experience and the RPC experience.

7.11.1 Clinical Reasoning

The CR of all subjects in this study were seen to correspond with what has been previously reported on directions of thinking for categories of hypotheses in the physiotherapy literature (Jones, 1995; Rivett & Higgs, 1997). The category of reassessment, identified by Rivett and Higgs (1997), however, did not seem to appear in the verbalisations of the subjects in this study. Figures 7.46-7.54 provide some examples of the data that was coded using the CR codes.

Figure 7.46 illustrates sources of symptoms which are structures or processes which are capable of emanating the client's signs, symptoms or dysfunctions.

-
- "I started thinking about maybe some form of inflammation causing him to wake up at night"
 - "I thought it might be good to check if there was any X-ray which might have shown some tearing of any tendons or any bony growth along the acromion or any bony growths causing impingement"
 - "Calcification, biceps tendonitis, impingement, that was the sort of thing that was going through my mind"
 - "I was trying to find out whether it was deep inside the capsule, whether it was superficial, maybe to do with muscle or ligament"
 - "I was thinking about the cervical and thoracic spine as well"
-

Figure 7.46: Source of Symptoms: Examples from Subjects

Figure 7.47 illustrates mechanisms of the client's signs and symptoms. These are the neural mechanisms which mediate the patient's signs and symptoms and include peripheral, central, sympathetic and affective components.

-
- “I wanted to make sure it wasn’t just muscle, just wanted to check muscle power mainly for the neuro type thing.”
 - “I did not think it was nerve related to anything in the vertebral column being impinged upon, any foramina closing or anything like that”
 - “Thinking maybe it could be referred pain from the cervical thoracic area”
 - “I thought it was maybe painful inhibition”
 - “I was thinking it was good that he had at that point no pins and needles and numbness because we could rule out worrying about any neuro tests”
-

Figure 7.47: Mechanism of Signs and Symptoms: Examples from Subjects

Figure 7.48 illustrates contributing factors which may predispose or may be associated with the development or maintenance of the patient’s problem. These can be environmental, behavioural, psychosocial, physical or biomechanical.

-
- “I was thinking the sleep position might give me a good idea as to the pathology”
 - “I was thinking there he had a heavy hammer, he was in a tight area and he was upside down, so I was thinking of impingement syndrome”
 - “When he said combing his hair the first thing I thought of was lifting the shoulder, like going into shoulder flexion above the level of the head”
 - “I detected that he had done something that he had not normally performed that could have brought on the irritability. Something, heavy object, overhead activity, prolonged, unaccustomed to it”
-

Figure 7.48: Contributing Factors: Examples from Subjects

Figure 7.49 illustrates naming or generating a hypothesis. In this category, the subject actually names a specific diagnostic hypothesis or syndrome. In the case of physiotherapy, this can also be a physical diagnosis.

-
- “I was just trying to rule out inflammation of the tendons”
 - “I was thinking mainly of say shoulder subluxation”
 - “I was just thinking fibrosis”
 - “I was thinking at the time it was a specific incident which happened therefore it was a tear of the supraspinatus tendon”
 - “These tests were coming out positive so I thought, biceps tendonitis”
-

Figure 7.49: Naming a Hypothesis: Examples from Subjects

Figure 7.50 illustrates precautions and contraindications to physical examination and/or treatment. In this case, subjects generally make note of having to moderate their physical examination, or avoid parts of it completely, because of the client's functional limitations and/or pain.

-
- “I wanted to find his end range ... I did not because of his pain, it made me apprehensive”
 - “I was thinking if I lift it up too fast and too hard I'm going to cause him pain, so I'm trying to ease into the position”
 - “I thought I want to make sure the pain goes away before I do the next movement.”
 - “Thinking here that I know the test is going to cause him pain and questioning the whole time of how much pain I should create with him.”
 - “Thinking about irritability levels just when he said 10 seconds, just trying to formulate how much we are going to be able to move it.”
-

Figure 7.50: Precautions and Contraindications: Examples from Subjects

Figure 7.51 illustrates minimising CR errors. In this category, the subject recognises the need to carry out certain tests or actions, or to ask specific questions, to minimise the possibility of a reasoning error.

-
- “He's already told me he's got no problems with his neck, I'll just check it anyway.”
 - “I thought okay, now we've got to go through clearing cervical”
 - “I thought, okay I'm gung ho again into this impingement thing, maybe I should back off and look at some other things rather than tunnel visioning into one thing.”
 - “I wanted to double check and make sure I cleared the acromio-clavicular joint”
 - “Weight loss, this bloke looks fine and healthy but I shouldn't make assumptions”
-

Figure 7.51: Minimising Clinical Reasoning Errors: Examples from Subjects

Figure 7.52 illustrates examples of management of the client that were verbalised by the subjects. In this category, indications of aims of management and/or citing treatment options for the client are made. This may include whether physiotherapy is indicated or appropriate for the client.

-
- “I wanted to tell him then, rest or do some passive movement to keep the length around there, ice, get the inflammation down and all that.”
 - “I thought if I could find out if he’s been having any treatment for it and if its helped, how that could be incorporated into the treatment programme for him.
 - “I was starting to think then the treatment of this patient would be critical because if he does not go back up on the shift that he’s meant to go up on for the next three weeks, that will be three weeks plus the week off. So that’s four weeks that we’ve got to treat him, and he’s going to lose all that income.”
 - “I was wondering whether to start on patient education now or leave it till later.”
 - “I was thinking here, having to do all this scapula control work, I thought he might think that it’s going to be quick fix treatment.”
-

Figure 7.52: Management of the Client: Examples from Subjects

Figure 7.53 illustrates the category of prognosis or outcome. This is an estimation of the extent to which the client’s problem appears amenable to physiotherapy. This may include the time frame that is expected for recovery and/or a prediction of the end result of treatment.

-
- “I was thinking we can do it in two weeks and get him back to work.”
 - “I was thinking am I being too conservative saying maximum eight weeks, is that really going to be better for him?”
 - “I thought three weeks, its going to be more than three weeks, it takes at least four or six weeks for any muscular kind of changes to happen.”
 - “I just had no idea how long it would take, how severe it was.”
 - “I was thinking at that stage, I wonder why he (the doctor) said for him to go to the physio, its obviously something that the doctor thinks he can’t deal with or its a specific rehab problem so that why he must have sent him here.”
-

Figure 7.53: Prognosis: Examples from Subjects

Figure 7.54 provides some examples of psychosocial inquiry made by the subjects. In this category, subjects note that they are directing their search towards gaining a better understanding of the client's psychosocial factors. These psychosocial factors are seen to possibly influence the assessment, management and/or prognosis of the case.

-
- "I was thinking to myself at this stage has he got a wife, a girlfriend, why isn't someone helping him, but maybe he lives alone."
 - "I started thinking that work is a concern for him in terms of finances, so I thought maybe there was some other alternatives at work that could get him back to work."
 - "When I said leisure time I was thinking of sporting activities and I wanted to pursue that later."
 - "Just thinking of how his injury has affected his work."
 - "I just wanted to explore his work situation just a little bit more."
-

Figure 7.54: Psychosocial Inquiry: Examples from Subjects

These examples illuminate the CR of the subjects in the IND and RPC groups in a very descriptive manner. The CR outcomes of this study, however, can also be examined in a quantitative way using clustering or counts.

Qualitative aspects of CR between the IND and RPC group were examined in two ways. The first method was to measure the frequency of each subject's verbalisations within each CR category. Mean values and ranges could then be calculated describing the breadth of reasoning that took place. For example, one subject in the IND group may have recalled three specific hypotheses, namely: impingement syndrome, bicipital tendonitis and rotator cuff tear. Another subject within the IND group may have recalled impingement syndrome and supraspinatus tendonitis only. The mean for this group, in terms of total number of hypotheses, therefore, is 2.5.

The second method is to look at absolute differences across the IND and RPC groups. Using the same example that was described in the previous paragraph, the

IND group would be seen to have generated four different hypotheses: impingement syndrome, bicipital tendonitis, rotator cuff tear and supraspinatus tendonitis.

Using the first method, differences in the IND and RPC group are presented in Table 7.25. These values represent the frequencies of the CR codes generated by subjects in the IND or RPC group. It is important to note that multiple iterations by a subject of the same thought were only counted once in the categories of: source of symptoms; mechanism of signs and symptoms; contributing factors; and naming a hypothesis. For example, if a subject said “I am thinking about impingement syndrome” five times over the course of a recall session, this hypothesis (impingement syndrome) was only counted once. This helped to control for situations where one subject may repeat the same hypothesis whereas another may only say it once or twice but still hold onto it as part of their reasoning framework.

The other categories: precautions and contra-indications; minimising reasoning error; management; prognosis; and psychosocial inquiry were counted as often as they appeared as each statement that is made within each of these categories is typically different from one another. For example, as subjects work through the SP encounter they may make several comments about how they must modify various aspects of the examination in light of the patient’s pain. These would all be counted as a separate thought about precautions and contra-indications. Interestingly, the means and ranges of scores within the RPC group, illustrated in Table 7.25, are slightly higher in comparison to the IND group. This suggests that there may be more free associative cognitive activity, with respect to the generation of ideas, thoughts and theory, in the RPC group.

Examining data across the two groups using the second method of analysis yields a rather different outcome. This is illustrated in Table 7.26. As can be seen from this table, there is little difference between the two groups with respect to the absolute numbers or types of verbalisations within each CR category, suggesting that the breadth of reasoning does not differ dramatically in either the IND or RPC group.

Table 7.25: Mean Scores and Range of Scores for Clinical Reasoning Coding Categories by IND and RPC Group

Category	IND Group n = 6		RPC Group n = 6	
	Mean	Range	Mean	Range
Source of Symptoms	16.5	13 - 26	21.0	17 - 27
Mechanism Signs & Symptoms	4.0	1 - 8	5.0	2 - 9
Contributing Factors	10.0	9 - 12	10.0	9 - 12
Naming a Hypothesis	3.7	2 - 5	4.3	3 - 5
Precautions & Contraindications	5.0	2 - 12	8.3	5 - 11
Minimising Reasoning Error	5.0	2 - 7	6.0	3 - 9
Management	1.2	0 - 4	5.0	3 - 7
Prognosis	2.5	0 - 5	4.0	1 - 8
Psychosocial Inquiry	3.3	1 - 7	8.7	8 - 10

Tables 7.27 - 7.30 describe the actual verbalised items within each category by IND and RPC group. These tables illustrate the similarities and differences in the breadth of thought between the two groups. Items which both the IND and RPC groups noted as part of their thinking are itemised under the ‘similar’ category. Items which either the IND or RPC group noted as part of their thinking are itemised under the ‘different’ category.

Table 7.26: Absolute Number of Items within a Coding Category: IND and RPC Group

Category	IND Group	RPC Group
Source of Symptoms	43	43
Mechanism of Signs and Symptoms	12	10
Contributing Factors	23	18
Naming a Hypothesis	9	8

Table 7.27: Actual Number of Items Within the Source of Symptom Category: IND and RPC Groups

IND Group Item Count = 43		RPC Group Item Count = 43	
Similar	Different	Similar	Different
• Supraspinatus Tendon/Tear	• Eccentric Loading	• Supraspinatus Tendon/Muscle	• Infraspinatus
• Impingement	• Glenohumeral Rhythm	• Impingement	• Teres Minor
• Cervical Spine	• Scapulo-humeral Rhythm	• Cervical Spine	• Subscapularis
• Boney-Growth, Osteophytes	• Intervertebral foramina	• Boney Growth	• Scalenes
• Inflammation	• Levator Scapula	• Inflammation	• First Rib
• Subacromial Space	• Serratus Anterior	• Sub-acromial space	• Labrum
• Acromio-clavicular Joint	• Wrist	• Acromioclavicular Joint	• Supraspinatus Fossa
• Rotator Cuff	• Swelling	• Rotator Cuff Tear/Rupture	• Back
• Deltoid	• Neural Tension	• Deltoid	• Forearm
• Biceps Tendon	• Cervical/Thoracic Posture	• Biceps Tendon	• Stiffness
• Elbow	• Thoracic Spine	• Elbow	• Scapula
• Bicipital Groove	• Lumbar Lordosis (posture)	• (Bicipital) Groove	• Scapulo-Thoracic Joint
• Stiffness	• Scoliosis	• Sterno-clavicular Joint	• Compression
• Acromion	• Cervical Level 7 - 8/T1	• Capsule	• Crepitus
• Sternoclavicular Joint	• Dura	• Glenohumeral Instability	• Dislocation
• Capsule	• Greater Tuberosity	• Pectoralis Minor	• Delayed Onset Muscle Soreness
• Joint Instability		• Trapezius	• Tumour
• Pectoralis Minor Tightness		• Pectoralis Major	• Cauda Equina
• Lower Trapezius		• Ligaments	• Bursa
• Middle Trapezius		• Transverse Humeral Ligament	
• Upper Trapezius		• Capsular ligaments	
• Pectoralis Major Tightness		• Tenderness	
• Ligament		• Scapular Winging	
• Transverse Humeral Ligament		• Vertebral Artery	
• Tenderness			
• Scapular Winging			
• Vertebral Artery			

Table 7.28: Actual Number of Items Within the Mechanism of Signs and Symptoms Category: IND & RPC Groups

IND Group Item Count = 12		RPC Group Item Count = 10	
Similar	Different	Similar	Different
• Pins and needles	• Muscle Power	• Pins and Needles	• Painful inhibition
• Numbness	• Nerve problems	• Numbness	• Sharp / Dull Pain
• Irritability	• Dura cord	• Irritability	• Cauda Equina
• Cervico-thoracic referred pain	• Neural Tissue Provocation Tests	• Cervico-thoracic referred pain	
• Constant / Periodic Pain	• Vertebral Artery	• Constant / Periodic Pain	
• Deep/Superficial Pain		• Deep/Superficial Pain	
• Nerve Root C5 Impingement		• Nerve Root Impingement	

**Table 7.29: Actual Number of Items Within the Contributing Factors
Category: IND and RPC Groups**

IND Group Item Count = 23		RPC Group Item Count = 18	
Similar	Different	Similar	Different
<ul style="list-style-type: none"> • Work related activity • Overhead activity • Abduction • Flexion • External Rotation • Hammer/Drill w. heavy object • Sporting activities • Tight pectoralis muscles • Functional activities • Muscle shortening/imbalance • Sleep positions • Fatigue • Unaccustomed activity • Previous pathology • Joint instability • Sustained positions 	<ul style="list-style-type: none"> • Scapular Control • Extension • Depression • Other treatment • Postural changes • Eccentric loading • Internal Rotation 	<ul style="list-style-type: none"> • Work related activity • Overhead activity • Abduction • Flexion • External Rotation • Hammer/Drill w. heavy object • Sporting activities • Tight pectoralis muscles • Functional Activities • Muscle tightness • Sleeping Position • Fatigue • Unaccustomed activity • Previous pathology • Muscle resistance (eg. conc/eccen) • Joint instability • Sustained positions 	<ul style="list-style-type: none"> • Lifting • Repetitive movement/trauma

**Table 7.30: Actual Number of Items Within the Naming a Hypothesis
Category: IND and RPC Groups**

IND Group Item Count = 9		RPC Group Item Count = 8	
Similar	Different	Similar	Different
<ul style="list-style-type: none"> • Rotator Cuff Tear • Impingement syndrome • Calcification within Tendon • Biceps Tendonitis • Tendonitis 	<ul style="list-style-type: none"> • Strain • Cancer • Shoulder subluxation • Supraspinatus tear 	<ul style="list-style-type: none"> • Rotator Cuff Tear • Impingement Syndrome • Calcification within Tendon • Bicipital Tendonitis • Supraspinatus Tendonitis • Tendonitis 	<ul style="list-style-type: none"> • Fibrosis • Bursitis

Subjects in both the IND and RPC group were seen to generate hypotheses early in their SP encounter, which is in keeping with what is reported in the literature. This is illustrated in Tables 7.31 and 7.32 respectively. The procedure for measuring the appearance of a hypothesis, in relation to the overall encounter, was determined as follows. First, the total number of lines in the complete transcript was calculated (column A). Second, the line number where the hypothesis appeared was determined (column B). Third, the line number of the actual hypothesis was divided by the total number of lines in the transcript, yielding a percentage score (column C). This

percentage score indicated at what point in the overall encounter, a hypothesis appeared. As can be seen in column C, most subjects recalled a hypothesis very early in the SP encounter. The majority of hypothesis, as indicated by their appearance in the text, became apparent within the first quarter of the verbal record.

Table 7.31: Appearance of Hypotheses in Relation to the Overall Encounter: IND Group

Subj.	Total Lines in Transcript	Hypothesis (Ho)	Line Location of Hypothesis	Percentage Value: B/A
	A		B	C
1: Low	491	Shoulder subluxation	15	3%
		Impingement	53	11%
		Supraspinatus impingement	124	25%
		Osteophyte formation	237	48%
		Supraspinatus Tear	307	63%
2: Low	609	Impingement syndrome	60	10%
		Calcification (tendon)	150	24%
		Biceps Tendonitis	150	24%
		Cancer	190	31%
		Supraspinatus tendonitis	492	81%
3: Low	427	Muscle impingement	30	7%
		Tendonitis	30	7%
		Rotator Cuff impingement	89	21%
		Supraspinatus (impingement)	209	49%
		Biceps (impingement)	209	49%
4: High	728	Impingement syndrome	259	36%
		Rotator cuff tear	333	46%
		Biceps Tendinitis	668	92%
5: High	649	Impingement	55	8%
		Biceps tendonitis	405	62%
6: High	366	Strain	20	5%
		Rotator cuff impingement	53	14%

Table 7.32: Appearance of Hypotheses in Relation to the Overall Encounter: RPC Group

Dyad	Total Lines in Transcript A	Hypothesis (Ho)	Line Location of Hypothesis B	Percentage Value: B/A C
1 High High	1297	Impingement	228	18%
		Supraspinatus Impingement	232	18%
		Tendonitis	389	30%
		Tear Supraspinatus Tendon	393	30%
		Rotator Cuff Tear	721	56%
2 High Low	1178	Fibrosis	298	26%
		Rotator Cuff Impingement	482	41%
		Calcific (boney) Tendinitis	483	41%
3 Low Low	859	Impingement	50	6%
		Bursitis	474	55%
		Bicipital Tendonitis	560	65%
		Supraspinatus Tendonitis	581	68%
		Inflammation of Tendons	682	79%

7.11.2 The Learning Experience

The second category of codes dealt with generic features of the learning experience that apply to both the IND and RPC groups. Figures 7.55 - 7.58 provide some examples of the textual data that was interpreted using the learning experience codes.

Figure 7.55 illustrates situations in which subjects realised that they were lacking in knowledge or uncertain about how to proceed.

-
- “That’s what I was thinking at that stage. I did not know exactly how to eliminate everything.”
 - “When he said that I had no idea what that drug was?”
 - “I did not know how to go about doing the proper orthopaedic tests”
 - “I had no idea what Feldene is.”
 - “I did not know whether to tell him I’m going to take my arms away or just to drop him....I did not know whether it would hurt him or what.”
-

Figure 7.55: Knowledge Gap: Examples from Subjects

Figure 7.56 illustrates situations in which subjects express anxiety or apprehension during the SP encounter.

-
- “I was just a bit anxious at that point.”
 - “I realised that perhaps I wasn’t as comfortable as I thought I would be in a normal situation so I started to feel the pressure a little more.”
 - “I was too anxious to try and clarify, because I know when I get anxious I stuff up what I’m saying.”
 - “I was really apprehensive about how much he said it was hurting.”
-

Figure 7.56: Acknowledging Anxiety: Examples from Subjects

Figure 7.57 illustrates situations in which subjects note that they are lacking in confidence.

-
- “The more I did the less I was confident with my diagnosis.”
 - “I’m not enjoying it now, I’m not enjoying myself because I think I do not feel confident anymore.”
 - “I remember thinking here, I just do not feel competent at all.”
-

Figure 7.57: Lacking Confidence: Examples from Subjects

Figure 7.58 illustrates situations in which the subjects comment upon the nature of the SP encounter and the fact that it is a simulation.

-
- “I was thinking to myself that he’s (the SP) blank or is it he hasn’t read his notes and he has forgotten that part of it or wasn’t told the bit of information.”
 - “I was confused because I knew I was supposed to treat it like a real patient but I didn’t know if I should look for more signs, or if this was real or that he was faking.”
 - “That was about the only time during the whole interview that I felt like he was trying to remember something that he’d been told to say, and I had an awareness that he was acting rather than being real.”
 - “I was thinking to myself this guy can’t fake his humeral head position, but I noticed that it was sitting forward, fairly far forward.”
 - “I was actually thinking he is not going to have these things because he has not actually got the problem but he did have some.”
-

Figure 7.58: Simulation Noted: Examples from Subjects

The frequency of verbalisations, by coding definition, are illustrated in Table 7.33. These frequencies are presented for both the IND and RPC Group. From a quantitative perspective it appears that subjects in the IND group expressed more concerns about gaps in their knowledge and reported more episodes of anxiety. Students in the RPC group more frequently noted that the SP encounter was a simulation.

Table 7.33: Learning Experience Codes: IND and RPC Groups

Category	IND	IND	RPC	RPC
	Group	Group	Group	Group
	Mean	Range	Mean	Range
Knowledge Gap	5.2	3-10	2.3	2-3
Acknowledging Anxiety	3.3	1-10	0.7	0-2
Lacking Confidence	1.2	0-3	0.7	0-2
Simulation Noted	2.7	0-9	3.7	2-6

7.11.3 The Reciprocal Peer Coaching Experience

The third coding category dealt with generic features of the reciprocal peer coaching experience. Figures 7.59 - 7.65 provide some examples of the textual data that was coded using the RPC codes.

One theme that became quite apparent in reading the recall data from the RPC group was role determination. Students recalled that at times they felt as if they were in the way, felt left out, had difficulty interjecting without feeling like they were interfering with the flow of the interview, or were not sure how to work as a dyad. In contrast, they also reported specific strategies for clarifying roles and for involving the other party. Examples of determining roles, whereby a subject comments about, or is sensitive to, the role of each party, are described in Figure 7.59.

-
- “I started thinking now that I had not really involved John at all and I did not really know how to handle that. I just remember thinking I have got to try and involve her more.”
 - “I started thinking I wish Mary would say at what degree she thinks the pain is coming on, because I am trying to record it and watch it at the same time and do everything.”
 - “I was thinking of taking more of an active part now because I just wanted to do some of it myself and get involved.”
 - “I was trying to get involved as well because I thought, like we had sort of divided it as I was going to do subjective and John was going to do objective, but I thought it would be better if we both got a bit more involved.”
 - “I started thinking that before we had gone in we had made up a list of who was going to ask what. I was thinking sort of where are we, what haven’t we asked, so we did not leave anything out and to to try and balance out who was doing the questioning.”
-

Figure 7.59: Determining Roles: Examples from Subjects

A second theme that emerged to a lesser extent was that of affirming action. Several students recalled noting to themselves that they had affirmed the actions of their partner whom they were observing. These reflections are illustrated in Figure 7.60.

-
- “I was thinking then it was good that she had chosen not to do it in 90 degrees because the 90 degree position was definitely going to cause pain and it was going to probably do more harm than good and at least in this position you could still try and work out to some extent what the structures were without actually hurting him.”
 - “I was glad she asked that because that was the next question I wanted to get on to.”
 - “Thinking along the same track, yes.”
-

Figure 7.60: Affirming Action: Examples from Subjects

A third theme, which is in contrast to that of affirming action, is critiquing action. Students appeared to make more mental reflections in this category than in the affirming category. Very few, however, actually verbalised their critique and coached their partner. What was apparent from these comments, however, was that peers have the capacity to provide instruction and immediate feedback (both confirming and developmental) to their fellow peers in many situations. For example, instructing them to carry out a test in a more biomechanically friendly way, altering

their hand position during a test, probing them to see if they can figure out a better way of asking a question. Figure 7.61 provides several examples of critiquing action.

-
- “I was thinking that John was hopefully looking at the scapula and where they were moving because really he should have done it the other way round, with his back to us, so we could see what was happening with the scapula.”
 - “I was just thinking oh why has he gone to the neck when he hasn’t got the shoulder pain yet?”
 - “At this stage I was hoping that Mary did not get too scientific with him because I thought that he did not need to know excessive detail about his shoulder.”
 - “When you started asking those questions I started thinking, I’m sure we have talked about this before, like we are starting to go over ground that we’ve already covered. I was thinking, why are we going back to that?”
 - “I remember thinking that what John said was a little bit wordy.”
-

Figure 7.61: Critiquing Action: Examples from Subjects

A fourth theme which emerged was that of seeking clarification. Subjects recalled noting to themselves that they wanted to obtain specific information during the SP encounter. However, because they were in the coaching role at that time, they found it difficult to get this piece of information because it meant intervening in the process and possibly breaking the train of thought or actions of their partner who was actively engaged with the client. Figure 7.62 provides some examples of seeking clarification.

-
- “I wasn’t sure what John was doing but I thought that’s all right when we’ve finished I’ll ask him exactly what he’s up to.”
 - “I found when John was palpating I was observing and not actually feeling it, I wasn’t too sure what he was feeling, where he was feeling ... I wanted to put my hands on the patient but because we had a lot of objective assessment I did not want to aggravate that.”
 - “When she asked is he working currently, I was hoping to try and get on to whether he was doing modified duties or not, but we did not get on to it.”
 - “I can remember when he said golf, I felt like asking more about his golf. How often does he play, how important is it to him, that sort of stuff.”
 - “I was thinking I wanted to explore the pain a little bit more, but Mary went on to asking about treatment.”
-

Figure 7.62: Seeking Clarification: Examples from Subjects

A fifth theme which emerged was intercepting a course of action positively (Figure 7.63). In some situations, the coach would actively intervene in the subjective or objective part of the SP encounter leading to a better outcome. Surprisingly, there were virtually no incidents of the opposite situation, save for a couple of comments, where intercepting a course of action turned out to be negative.

-
- “I was actually quite happy that Mary asked that question because it was actually something that I thought was important for the duration, and I had definitely not thought to ask that question.”
 - “I became aware of the fact that when Mary said, “do you need a rest”, then I kind of thought to myself, hang on, I can’t just plow through all these tests because this guy has got pain.”
 - “I was thinking that it was good that John cleared it up, because I mean it’s hard with the steroid question.”
 - “I was trying to think of some other way to ask it, and that’s when John came in.”
 - “I had just assumed he was hammering normally, then when John asked him how he was hammering and the patient was explaining it...I was thinking oh yes, he could have been hammering up like that on the ceiling...it’s good that John asked that and got the truth.”
-

Figure 7.63: Intercepting a Course of Action Positively: Examples from Subjects

A sixth theme which was clearly evident was that of support. Subjects frequently noted the presence of their peer and the emotional and practical support that they received from them. Figure 7.64 provides several examples of reflections in this category.

-
- “I was actually thinking then I hope that you take over that little bit because I did not know what it was and maybe you ‘d get some information that I would not be able to get because I did not know what it was.”
 - “Yes I was thinking the same thing, I was thinking Feldene, oh that’s anti-inflammatory, looking to John for reassurance.”
 - “I was thinking is this just me?, better get John to have a go.”
 - “I sort of looked to John whether we should do it or not.”
 - “I ‘d run out of things to say. I looked over to Mary and thought, Mary you can carry on now I’ll just sit back and think for a while.”
-

Figure 7.64: Support: Examples from Subjects

A seventh theme was competition. This only manifested in one of the RPC groups - the high:high group. Competitive comments, however, were rather frequent and were indicative of the problems this particular RPC group were having. Figure 7.65 provides several examples of competitive reflections made by the students in this dyad.

-
- “I did not feel in control of the situation at all so I felt like I was just waiting. By this stage I was feeling I just wanted to do that myself without someone else there.”
 - “I felt it just all wasn’t clear in my head at all because I did not go about it my way.”
 - I was just thinking I’ll try and keep out of the road because I’d had my turn with the subjective and I knew Mary would be wanting to do it the way that she wanted to do it, she might find that I was a bit distracting and sort of getting in the road.”
 - “We just did not communicate with each other at all.”
-

Figure 7.65: Competition: Examples from Subjects

The frequency of verbalisations in the RPC coding categories are illustrated in Table 7.34. It is interesting to note the outcomes from the high:high group. As was mentioned earlier, this group was quite competitive and had difficulty working together collaboratively. One can see from the last column that several verbalisations fell into the competition category. In light of this competition, there appeared to be a greater frequency of comments concerning roles (DetRol). Further, there were also a lot of actions which were critiqued by the other party (CriAct) and expressions of wanting to have information clarified (SkClar). These outcomes were less prominent in the low:low and high:low pairings. Because these two groups appeared to be less competitive, one can observe a greater incidence of intercepting action positively (IntAct+) and support in these two groups.

Table 7.34: Reciprocal Peer Coaching Experience Codes: IND and RPC Groups

RPC Group	DetRol	AffAct	CriAct	SkClar	IntAct+	Support	Comp
Low:Low	9	2	5	5	10	4	0
High:Low	12	1	2	7	1	10	0

High:High 24 1 15 11 3 6 7

Key: detrol=determining roles; affact=affirming action; criact=critiquing action; skclar=seeks clarification; intact+=intercepts action positively; comp=competition

7.11.4 The Reciprocal Peer Coaching Experience - Student Evaluation

All six subjects in the RPC group were also asked a series of questions after the stimulated recall session which asked them to reflect upon their experiences as a coaching dyad. For the most part, the students' comments were very positive and are summarised in the following paragraphs.

Question One: What did you like about the reciprocal peer coaching experience?

Students saw the reciprocal peer coaching experience as an opportunity to jointly problem solve, obtain support, and to facilitate learning and performance. While it was recognised that the students needed further practice to refine their collaborative skills, they recognised the value of working together to meet a common goal. The following quotations illustrate these features.

“I liked the interacting and having someone there...but its not something I think we're good at at the moment. I can see so much value that you could get out of it.”

“It was good because we had someone to practice with and we could talk through our problems and it was just back up, like if you felt you were going nowhere you could look to John or whatever and say how do you feel?”

“We sort of said okay I'll do this section generally, you do that section, we'll cover up each other, back each other up if we miss anything. And the key thing is we said we'll pause so that the other person can interject. And I felt that we were able to do that quite well without the other person thinking, oh they're having a go at me, I'm being stupid. It did not feel that one was trying to know more than the other, it felt really good. It enabled us to cover everything, it's so easy on your own to just fly through it and forget big chunks.”

“It was an equal mark, we both felt that we were more or less like a team, you were working to a joint goal, it wasn't I'm going to make up my little mind and he's going to make up his little mind and we'll be competitive in the middle. It was we both needed the information to get where we were going.”

Question Two: What did not you like about the reciprocal peer coaching experience?

There were also specific aspects of the RPC experience that frustrated or displeased the students. One of them was dealing with the fact that your peer may have a different approach or tact which did not necessarily coincide with the other peer's preferred mode of operating. Wanting to support their peer, but not feeling comfortable doing this in front of the patient, was also a frustration that was noted by the students. The following quotations illustrate these features.

“We had very different styles I think of going about it. There was the frustration with me wanting to get the information my way and Mary had her way and therefore I couldn't get the same type of information out of it that Mary did.”

“Sometimes John would go off on his own tangent and I wanted to do something else.”

“Conferring about things in front of the patient...you can't really do that, whereas if your by yourself you just go, okay this, this, this, maybe this.”

“How do you confer openly without making the patient feel like they're having a secret meeting.”

“I did not know how to involve Mary in the objective part of the assessment.”

“It was an issue for me to be evaluated as a pair”

Question Three: Were there any aspects of the reciprocal peer coaching experience that facilitated your clinical reasoning or performance?

Again, the subjects found the process of working with a peer conducive to the development of their clinical reasoning. This was particularly apparent in the 15 minute discussion period after the SP encounter. This discussion was useful for the subjects to coalesce their ideas and views about the patient's signs, symptoms, diagnosis and management plans. Gaps in knowledge, assumptions, mistakes, errors in reasoning could all be exposed and reviewed leading to a more solid conclusion about the case. The following examples illustrate these features.

“It was good because you had two minds thinking about the different things. Like when we were talking afterwards in the 15 minutes about the diagnosis. We could put our own, our two ideas together.

“...part of it that was possibly good was at the very end when we had 15 minutes to talk about it.”

“That 15 minute interval was more clarification of everything that we'd gone through to make sure that we both understood the same things - cause we had done different parts of the examination.”

“One person would mention something that the other person possibly had not thought of or had thought of but did not remember it at the time. So it was good to confirm and recap.”

“When we did the apprehension type tests, we started off with one thing and then one thing led to three things and that then came to a better diagnosis I think than what the individual test would have done...it sort of facilitated some of the tests to be given a more accurate result.”

“I think John sort of tended to jump to diagnosis faster than I did, and then I sort of said okay what about this, what about that, and backed up. And then he would justify that and sort of say okay maybe its not that, but then there’s this and that which leads to that diagnosis. So, therefore, as a team we sort of came to a similar diagnosis with all the information there.”

Question Four: Where there any aspects of the reciprocal peer coaching experience that hindered your clinical reasoning or performance?

For the most part, frustrations about the RPC experience and its negative effects on CR came from the high:high group. This group was quite competitive in its approach to the SP encounter and had difficulty working collaboratively. Because of this inability to work collaboratively, information and task sharing decreased and resulted in the other partner having an information gap. This in turn made the reasoning process more difficult because of incomplete information. The following quotation illustrates this point.

“I did not feel part of the whole objective, so I sort of felt that I got lost in between subjective and objective. I did not feel I was really able to make good decisions because I had to rely completely on Mary and I couldn’t add anything to what she said other than what I’d got out of the subjective. So if she told me that it’s the supraspinatus that the patient’s feeling pain in, then I had to agree with her because I couldn’t say no.

Question Five: In what ways were you able to support your partner in preparation for the patient encounter?

Of the two RPC teams that were able to work collaboratively, it was clear that they had put some preparation into actually sorting out how they would work together as a team. The competitive group did not discuss strategies for team work. They only reviewed each others’ plans for the assessment to make sure no information was going to be overlooked. The following quotations illustrate the strategies that the more cooperative coaching dyads employed.

“We sort of told each other what we were doing in the last couple of days. We had a couple of meetings and I wanted to have a couple of videos and have a look at them. Then we came in here this morning and we both found articles from separate areas, and pulled it together and had quite a good little knowledge basically.”

“We were able to arrange times and practice, go through some theory, pop questions at each other, and generally discuss what the options were.”

Question Six: What suggestions can you make to improve the RPC process?

The importance of practice was again identified as a key strategy for ensuring that the group was able to work together effectively. Learning from experience, having good interpersonal communication skills, being comfortable receiving developmental feedback and having some flexibility appear to be important facets of a successful working relationship. The following quotations illustrate these features.

“Just experience it with each other, interacting, getting each other working a lot more together, it does not even have to be related to the problem. Just make sure you understand each other and where you’re coming from and make sure that your communication between each other is perfect so that one person does not misinterpret another thing that you’re saying or that you feel confident that the other person can criticise you objectively and you’ll be completely fine with that. That’s really important because otherwise you may interpret their criticism wrongly and the relationship would break down.”

“Just allow the other one to speak up and not to come in there with your own fixed ideas, be flexible.”

Question Seven: If you could go back and repeat this experience, what might you have done differently?

All of the three groups were able to identify things that they would have done differently. Given that this was their first experience carrying out a complete patient examination as a peer coaching dyad, they were aware of the fact that they needed to continue to develop their team learning skills. Practice was seen as a key feature of improving the dynamics of the RPC experience. Developing strategies for integrating the other partner actively in the assessment process was another aspect needing improvement. Communication and feedback systems that would work in front of the patient also needed to be developed. The following quotations illustrate these features further.

“Work together a bit more”

“Practise talking to each other and practise involving each other. I think that is where we lost so much... both of us could have learnt so much from each other.”

“We did not have a way of communicating to each other that could be seen as a criticism in front of the patient.”

“Try to integrate a bit more.”

“Know where your concluding points are”

“Do a complete mock up run of it through on our own before we come in. I felt we overlapped and double backed a few times. A few more times than what we’d actually planned on doing.

Question Eight: If you had a choice as part of your learning, would you prefer to always see patients individually, with a peer coach or as a combination of both? Why?

All subjects felt that seeing patients individually and in a peer coaching system had advantages and disadvantages. Subjects recognised that they would need to work autonomously as a health professional when they graduated. Hence, they needed to be able to clinically reason through a case independently. However, they also recognised that they were developing their skills as a practitioner and there was great value in collaborating with another individual with respect to testing out ideas and conclusions. The following quotations reflect these feelings.

“Probably a combination of both because in the end you are going to have to work on your own.”

“A combination of both, although I think maybe if you go in as a single person but you have a peer coach in the background and they are there to see what sort of results you get.”

“I would like to see a patient individually first, then go into some peer stuff, and then go back and do some individual work. I think that way I’d get a lot out of it because I’d go to my peer with my own prototypes and I could help and add things and she could add things, and we’d come up with something that was more complex.”

“I would like to go to a placement where you do things individually...so you have a picture in your mind... then get together with someone else and go through it, and then I think I’d be much more open to learning another way of doing it.”

The outcome of the RPC experience was positive when one considers the students’ comments globally. What is clear from the students’ feedback is that they need to be adequately prepared for this type of learning model as it is quite different from the individual model of learning. Practice is also something that will enable students to become more efficient at learning in teams. Through practice and preparation, it would seem that highly effective learning systems can be developed which may be transferable to the clinical setting.

Chapter 8: Discussion

8.1 Summary of Findings

8.1.1 Research Objective 1:

To determine differences in undergraduate physiotherapy student performance, from the perspective of a patient encounter, across the IND and RPC learning models.

For the most part, students in the reciprocal peer coaching (RPC) group outperformed their peers in the individualistic (IND) group in the areas of history taking, physical examination, and communication. Interestingly, this higher performance in the psychomotor domain was also much more homogenous than the scores of the IND group: as evident in the narrower standard deviations of the RPC group. This suggests that students in the RPC group, as a whole, did consistently better than their peers in the IND group.

The performance standards that were set by the panel for the checklists also appeared to be realistic. Overall performance on the combined history and physical examination checklists for the IND and RPC group was 62.5 and 70.0 per cent respectively. These are average scores and suggest that the simulated patient (SP) encounter and associated checklists were appropriately designed and benchmarked for pre-clinical physiotherapy students. By encouraging panel members to take into consideration the competency levels of the students, and to consider both actual and evidence based practice to set standards, a reasonably challenging and valid testing situation was created.

8.1.2 Research Question 1:

How does history taking skill differ in the IND vs. RPC Group?

On a specific level, differences in history taking skills were practically significant; there was a moderately positive effect size, with the RPC group indicating the best skills (this difference, however, was not statistically significant). Differences in history taking ability may not be influenced to the same degree as other competencies

as both groups used a standardised menu format for their interview. Questions on these menu formats tend to be relatively standardised in physiotherapy and all one has to do is go through the checklist. The fact that the RPC group still did better on a practical level, however, is interesting. What appears to be the case, as became evident in the qualitative interviews after the encounter, was that peers could monitor whether questions were asked/not asked, could bring the interview back on track, or probe the patient further if responses were too ambiguous. This improved the quality and thoroughness of the history taking encounter.

8.1.3 Research Question 2:

How does physical examination skill differ in the IND vs. RPC Group?

Strong positive effect sizes were evident in favour of the RPC group. Differences in physical examination scores were statistically significant with the RPC group outperforming the IND group. Again, information from the qualitative interviews suggest that peers could monitor the efficacy and thoroughness of testing, corroborate findings, and encourage and support their partner to carry out further testing. Many of the students in the IND group did not extend the SP as far as the students in the RPC group. The SP's pain levels, which were moderate but unaffected by changes in range of motion or movement, were of great concern to all students. However, students in the RPC group appeared to be able to extend the patient further because they were supported and encouraged by their peers. Students in the IND group would often not extend the patient once he reported pain. Because the students in the RPC group extended the patient further, they collected higher quality information. The higher quality output and thoroughness by the peer group would theoretically increase this group's self-efficacy, thus enhancing their belief in their ability to carry out these skills in future practice (Bandura, 1977; Bandura, 1997; Perry, 1988).

8.1.4 Research Question 3:

How does interviewing skill differ in the IND vs. RPC Group?

Students in the RPC group also presented with higher scores on communication skills, as measured by the Arizona clinical interviewing rating scale. The RPC group

outperformed the IND group with a strong positive effect size being evident. The higher performance of the RPC group was also statistically significant. The reasons for this improved performance may be due to the extra attention the SP receives in a RPC situation. For example, while one of the students was performing a test and deep in concentration, the other student could keep the SP engaged in a conversation. Further, towards the end of the encounter, the SP was trained to ask the students when he could go back to work because he was quite anxious about not earning any money. When students were vague or non-committal about the prognosis, the SP was trained to become rather persistent about getting a time frame for treatment. In this difficult situation, the peers could share the ‘heat’ of the moment and could intercept on behalf of one another. This would settle the patient and resulted in a more positive communication outcome. Students in the IND group had to fend for themselves.

8.1.5 Research Question 4:

Is there a difference in time it takes to complete the task, thoroughness, and efficiency, across the IND and RPC Groups?

There were interesting differences in the time it took to complete the SP encounter. There were also some noteworthy differences in terms of thoroughness and efficiency across the IND and RPC groups. Students in the RPC group basically took longer to complete the overall SP encounter. One reason stems from the interaction that occurs during the encounter. Students in the RPC group would often ask for their peers advice, invite them to ask other questions, or request them to corroborate a physical finding. This understandably added to the length of the SP encounter. Another reason the RPC group took longer to complete the SP encounter was because they were more thorough. The students in the RPC group were significantly more thorough on all aspects of the SP encounter with this being particularly noteworthy on the physical examination. There was also narrower variance in the thoroughness scores of the RPC group demonstrating that there was more homogeneity in the performance of this group.

It is important to note that thoroughness is an index of how well the student(s) were able to complete the required tasks on the history and physical examination checklists. A competency strategy was necessary, requiring students to strategically

question and examine the simulated patient using a comprehensive set of data inquiry methods. This is not to say, however, that both groups did not make errors or carry out unjustified inquiries. These were not measured in this particular study. It is safe to posit, however, that some students may have used very rudimentary forward reasoning skills in both the IND and RPC group. Clearly, some students did have previous experience with patients with shoulder pathology and they may have ‘skipped’ over some of the required tests and questions on the physical examination and history checklists. So, while in this particular study, the implementation of more history questions and physical examination manouevers would likely lead to a higher score, in some cases, students may have been caught up in the application of a performance strategy rather than a competence strategy. Given that both groups were relatively equal in background experience, one would expect that this forward reasoning effect would be balanced across the two groups. Hence, the fact that students in the RPC did better than their peers in the IND group, suggests that this outcome is genuine and not due to a forward reasoning effect.

Both the IND and RPC groups were equally efficient, but not necessarily equally effective. Given the thoroughness of each groups’ intervention, divided by the actual time it took to complete the encounter, both groups’ efficiency ratios ended up being mathematically equivalent. For example, the IND group was less thorough and took less time to complete the SP encounter whereas the RPC group was more thorough and took more time to complete the SP encounter. Given these ratios, one can see how the efficiency ratios would be similar. In the end, however, the RPC group was clearly advantaged in that it was efficient but also more effective than the IND group. Although it took the RPC group longer to complete the overall SP encounter, the RPC group’s data collection was much more thorough, resulting in a much more effective encounter. This effectiveness was advantageous as one could argue that it led to the RPC group’s greater PEQ scores. This would suggest a deeper level of learning or reasoning about the case.

In conclusion, the RPC model appeared to facilitate higher levels of performance in the psychomotor domain. This is in keeping with the literature which purports that peer assisted learning (PAL) techniques can be used to promote achievement. For

novice physiotherapists, therefore, the RPC model is a useful adjunct teaching strategy for developing practice skills.

8.1.6 Research Objective 2:

To determine differences in clinical reasoning and problem solving across the IND and RPC learning models.

8.1.7 Research Question 5:

Are there differences in clinical reasoning across the IND and RPC group from the perspective of diagnostic skill, ability to identify management options for the client, and skill in identifying the key features of the case?

The higher levels of achievement seen in the RPC group were also evident in the post-encounter clinical reasoning (CR) questionnaire. This higher performance was statistically significant. Moderate to strong positive effect sizes were also seen in favour of the RPC group on most sub-sections of the PEQ with the exception of the history key features items. This latter section was relatively similar across the two groups. Again, this being largely due to the relatively standardised menu formats used in physiotherapy to collect patient history data. A statistically significant difference was evident, however, on the patient management question with the RPC group obtaining a higher score.

Reasons for the RPC group's higher achievement on the overall PEQ likely stems from the collaboration that takes place throughout the encounter and from the discussion that follows the encounter. During the 15 minute interval between the SP encounter and the PEQ, students in the RPC group could discuss their CR and ideas

on patient management. This brainstorming provided a broader perspective on the patient which became evident in the PEQ scores.

It was interesting, however, that there were small effect sizes for the subsections of the PEQ other than for 'patient management'. One possible reason for not seeing a greater divide on the PEQ scores across the two groups is likely due to the employment of biomedical knowledge. Most students in the IND and RPC group were able to identify the SP's main diagnoses or problems. With these pieces of information, students would be able to access a store of biomedical knowledge on these problems. Rotator cuff pathology and shoulder injuries, for example, are well covered in the undergraduate curriculum. From this underlying biomedical knowledge, students could, theoretically, extrapolate pertinent history and physical examination key features from the multiple choice key features test - even if they had not actually performed those procedures during the SP encounter.

To illustrate this point further, and to demonstrate how it may influence outcome scores between the IND and RPC group, the following example is provided. Patients with rotator cuff pathology generally find sleeping on the affected shoulder painful as the pressure increases the pain. This is a particularly noteworthy or key feature of rotator cuff tendinitis and is a symptom that is commonly reported by patients with this pathology. A student who has determined that there is a rotator cuff tendinitis from their SP encounter will obviously see this sign as a key feature of this pathology. Whether or not the student actually asked the patient about sleeping positions and pain during the encounter is irrelevant because they can apply their biomedical knowledge to answer the question having concluded the correct diagnosis.

This reliance on biomedical knowledge to solve aspects of the PEQ is evident in the weak but statistically significant correlation between the SP encounter and the post-encounter CR test ($r=.298$). There was a weak but statistically significant correlation between the combined history and physical examination checklist score and the overall post-encounter CR score. This suggests some positive association between these factors. That is, a thorough history and physical examination correlated positively with a good PEQ outcome score. However, as mentioned earlier, the fact

that students could rely on other sources of biomedical knowledge to work through parts of the PEQ tended to weaken the final correlation.

A particularly noteworthy outcome of the study was related to the performance of the low achieving students on the post-encounter CR questionnaire. Low achieving students in the RPC group outperformed low achieving students in the IND group. The use of PAL to facilitate the performance of low achieving students has been identified in the literature and this study certainly supports this finding (Chapman, 1998; Smith, Hinckley, & Volk, 1991). Low achieving students in the RPC group outperformed the low achieving students in the IND group on all measures of the post-encounter CR questionnaire with strong positive effect sizes being evident. Statistically significant differences in scores were also seen in the patient management question and overall PEQ for low achieving students within the RPC group. Since all students, regardless of their group assignment, had to write the PEQ independently, these results suggest that the achievement of low achieving students was augmented by peer assisted learning.

High achieving students, in contrast, did not really differ across the IND or RPC groups on the PEQ except in the patient management question. Again, strong practical significance, which was also statistically significant, was seen in this category with the high achieving students in the RPC group outperforming their counterparts in the IND group. The collaboration during the SP encounter and the discussion that followed appeared to have been a key factor in improving the performance of the high students in the RPC group.

The consistently higher performance on the 'patient management' question in both the low and high achieving students in the RPC group suggests that the brainstorming of ideas helped students to put together a comprehensive management plan for the client.

In conclusion, the RPC approach was seen to be a powerful force in facilitating the CR skill of students. Through PAL, students in the RPC group were able to develop a much deeper understanding of the client.

With respect to actual reasoning processes employed by the physiotherapy students in both the IND and RPC groups, they all appeared to employ CR categories that have

been described in the physiotherapy and medical education literature. For example, evidence of verbalisations could be applied to most physiotherapy CR categories. Specifically: source of symptoms, mechanisms of signs and symptoms; contributing factors; naming a hypothesis, precautions and contraindications, minimizing CR errors; management of the client; prognosis and social inquiry (Jones, 1995). However, very few students used the category of reassessment that has been identified in the physiotherapy CR literature (Rivett & Higgs, 1997). For example, only one student in the IND group verbalised two thoughts that could be categorised as reassessment. Only one RPC group verbalised a thought that could be considered reassessment. One reason for this may stem from the fact that this SP encounter was a singular experience. The need to identify information that could be used in reassessment was not critical in this particular situation as it was a one off experience.

The RPC group also appeared to verbalise more thoughts in the categories of: sources of symptoms; hypotheses; precautions and contraindications; management ideas; prognoses; and social inquiries (Table 7.25). This may suggest that there was more free associative cognitive thinking going on in the RPC group. Further, the rather higher frequency of thoughts in the categories of social inquiry and management may indicate why the RPC group did so much better on the PEQ, in particular, the management question. With greater inquiry into the social aspects of the client's case and a better understanding of his physical limitations and prognosis, students in the RPC group had a deeper understanding of where the patient was coming from and what needed to be achieved in treatment. This deeper understanding of the patient's perspective relates to the narrative, pragmatic and predictive forms of reasoning that have been described in the CR literature (Benner, Tanner, & Chesla, 1992; Edwards, Jones, Carr, & Jensen, 1998; Fleming, 1991; Hagedorn, 1996; Mattingly, 1991; Schell & Cervero, 1993). These qualitative forms of reasoning appeared to be more frequent in the RPC group. There were also more iterations of precautions and contraindications in the RPC group which meant that there was more active thinking going on about how to moderate the physical examination, which may have then fueled this group's more thorough and comprehensive approach to patient assessment.

Looking at the mean numbers of verbalisations in each clinical reasoning category is, however, only one way of examining the qualitative data in this study. Looking at absolute differences between the IND and RPC groups is another way of examining the data (Table 7.26). By counting the presence of each specific CR label only once, one finds that there is little difference between the IND and RPC groups in terms of the content used to work through the case. Examining the data in this manner suggests that the CR process is not necessarily advantaged or disadvantaged when employing an IND or RPC learning model from the perspective of breadth in thinking.

The IND and RPC groups were also found to have employed hypothetico-deductive reasoning. Both groups were found to have generated early hypotheses (Figures 7.31 and 7.32). For example, when examining the recall data of the students, evidence of early hypothesis generation was evident in the data for both groups with hypotheses emerging early in the overall verbal record. The students in both groups were also seen to continue with their patient inquiry even though they expressed certainty about the patient's hypothesis. Part of this may stem from their understanding that they needed to perform a comprehensive examination. However, this continuation also appeared to be used to minimize reliance on confirmatory evidence, to minimize reasoning error and to build rapport with the client. This resulted in the students having greater confidence in their hypotheses and assured them that they had not missed any significant cues. Evidence of this error minimisation strategy can be seen in the students' comments in Figure 7.46.

The students in both groups were also quite preoccupied with data gathering, following their template of questions and collecting as much information as possible. This is quite different from experts who appear to be much more strategic about their questioning, responding predominately to critical cues. Evidence for this novice tendency to be highly focused on the process of data gathering has been reported in the literature and is certainly corroborated in this study (Embrey, Guthrie, White, & Dietz, 1996; Jensen, Shepard, Gwyer, & Hack, 1992; Jensen, Shepard, & Hack, 1990; Thomas-Edding, 1987).

8.1.8 Research Question 6:

Are there differences within the IND and RPC groups in terms of their history, physical examination and interviewing skills?

Intragroup Differences - RPC Group

There were three dyads within the RPC group: low:low; high:low; and high:high achievers. Given the small size of each group (n=7) it is difficult to ascertain whether any specific combination yielded any particular advantages. Further, the narrow range of spread in the academic grades of the low and high achieving students (Figure 6.2) may make the distinction between high and low achievers less than practical. For example, 56 per cent of students in this study had mean grades between 65 and 77 per cent for orthopaedic science. To illustrate how this distinction between high and low achieving students may be less than practical, the following explanation is provided. Students who enter the School of Physiotherapy program typically have very high academic grades and are representative of upper level academic achievement percentiles in the general student population. Once enrolled in the School of Physiotherapy, however, academic grades are normatively referenced (scaled) around a mean of 70 with a standard deviation ranging anywhere from five to twelve. Hence, the distinction between low and high achievers may be artificial given the actual potential of these students vis a vis the general student population and the somewhat artificial normative referencing system that takes place within the School of Physiotherapy.

Keeping this explanation in mind, there were still some differences across the three dyadic structures. For example, the high:low student group completed the physical examination in less time than did the other two groups. The same group also used less time for the overall encounter in comparison to the the other two groups. In light of this performance, the high:low group turned out to be the most efficient because they were also relatively thorough. It is difficult to ascertain why this particular grouping was so much faster than the other two groups although the literature on cooperative learning suggests that heterogeneous groupings appear to provide the best learning and performance outcomes (Greenwood, Carta, & Kamps, 1990).

Surprisingly, the low:low group outperformed the other two dyads on the history key feature questions on the PEQ. This strong practical significance was also statistically significant. The low:low group also outperformed the high:high group on the overall PEQ. This strong practical difference was also statistically significant. Although this seems strange it is possible that the low:low group, recognising their weaknesses, put a lot of extra effort into preparing for the experience. Alternatively, high:high groups may have been too competitive which interfered with the group's achievement. There is certainly some evidence for this in the qualitative reports of the high:high group. Lastly, it may merely be an artifact given the narrow academic differences that likely exist between the so-called low and high achieving students.

Other than these minor contrasts in the performance of the three groups, these rather indistinctive differences parallel findings reported in the literature (Riggio, Whatley, & Neale, 1994). Riggio et al. (1994) did not find any statistically significant differences in cognitive gains between high:low, low:low, high:high, or medium:medium groups in an undergraduate psychology unit. They argue that in a reciprocal peer tutoring situation, both students must be accountable and support one another. This requires that the students put more effort into understanding and preparing material for the tutoring session. Hence, Riggio et al. (1994) feel this heightened cognitive activity is pervasive across all groups, regardless of the actual combination, and thus washes out differences related to previous academic achievement.

What appears to emanate from these specific dyadic groupings is that working with a peer offers many learning advantages. As a learning model it is certainly no worse than the individual learning model. Further, it is very possible that low achievers can increase their individual achievement when paired with a peer. This has important ramifications for improving the clinical performance and reasoning skill of the novice practitioner. Whether low achieving students should be paired with another low achieving student or a high achieving student, however, is difficult to ascertain from this study. Sample sizes were too small to conduct an analysis at this level. One must also be careful about relying on academic marks as the sole predictor of an effective or ineffective learning dyad. It is perhaps more realistic to adopt the views of Goldschmid and Goldschmid (1976) who stated that there are so many complex

factors at play in a learning dyad that it is difficult to ascertain with any certainty the best pairing arrangement.

Intragroup Differences - IND Group

Unlike the RPC group, there were noteworthy differences in performance within the IND group. High achieving students basically outperformed the low achieving students on all outcome measures with moderate to strong positive effect sizes being evident. With this sort of outcome, one cannot help but assume that the split in the two groups using academic grades as the criteria had some credence. Unless of course, when assigns students to work together, this difference in academic potential disappears because of the joint support that takes place during learning. Nonetheless, the fact that the performance of low and high achieving students was quite distinctive in the IND learning group is interesting given that this distinction was not as evident in the RPC group.

With respect to the IND group, high achieving students did particularly well on the physical examination and on the overall SP encounter with statistically significant higher scores than their low achieving counterparts. The more thorough approach by the high achieving students on the SP encounter also took a statistically significant longer time than the low achieving students. Again, this is in keeping with the literature on expert and novice performance in physiotherapy and the results of the pilot study for this experiment. Experts, and the physiotherapists who participated in the pilot study, generally take longer to complete tasks as they spend more time with their clients in trying to understand the holistic perspective of the case. High achieving students in this particular study were found to have followed this example.

8.1.9 Research Objective 3:

To determine differences in the affective domain of clinical competence across the IND and RPC learning models.

8.1.10 Research Question 7:

How is anxiety and confidence influenced by the IND and RPC learning models?

Anxiety

Lack of social support and social isolation have been reported in the literature as factors which increase academic distress and dissatisfaction with learning (Fantuzzo et al., 1989). Peer assisted learning techniques have been used as a strategy to mitigate this psychological stress and anxiety. The results of this study provide evidence in support of the RPC method as a means of reducing anxiety levels in students. Whether this anxiety is test related, related to the actual clinical task and subsequent reasoning test, or anxiety from having to engage in a new and unfamiliar task was not directly addressed in this study. While both groups became less anxious as they progressed through the SP encounter it was the RPC group that became significantly less anxious. In other words, reduction in anxiety levels was greatest for the RPC group. This reduction in anxiety may have been due to the fact that the encounter was coming to an end. Alternatively, the greater reduction in anxiety seen in the RPC group, may genuinely come from the peer support inherent in this model. The importance of this support in the RPC group is evident in the qualitative data that was collected after the SP encounter (Figure 7.59). Students verbalised their appreciation of the peer support that the RPC model provided. It appears, this support helped to reduce anxiety levels and made the learning situation less stressful for the students.

Confidence

Interestingly, subjects in both the IND and RPC group became less confident as they worked through the SP encounter. This is in contrast to reports in the literature that clinical experiences and the OSCE generally increase confidence in students (Ytterberg et al., 1998). Neither of these downward trends, however, was practically significant. The reasons for this drop in confidence are not known. However, one may speculate that the inexperience of the students was a major factor in fueling their lack of confidence. For example, integrating biomedical knowledge with clinical experiences is a daunting task. Suddenly realising that your biomedical knowledge needs to be reconsidered or re-structured in light of a patient's physical presentation can be quite a challenge. For most of these students, this was the first time they had been given the responsibility to put their biomedical knowledge into practice. Qualitative comments made by the students after the SP encounter suggest that this

lack of experience was a major factor in influencing their confidence levels (Figure 7.52).

Research Question 8:

What experiences do students report when engaged in an IND vs. RPC learning model?

The Learning Experience

In terms of the actual learning experience, students in the IND group appeared to verbalise more concerns about gaps in their knowledge base and the presence of anxiety (Tables 7.50 - 7.51). This certainly supports the quantitative data of this study that demonstrated that the RPC group had a statistically significant decrease in their anxiety during the SP encounter. It also provides further support for the literature that argues that PAL techniques can be used to minimise learner anxiety.

The Reciprocal Peer Coaching Experience

Verbalisations from the students in the RPC group on the peer coaching experience yielded interesting information about this learning model. Role determination appeared to be a rather significant issue for the students. Wagstaff (1989) has noted that there may be confusion about roles if students are encouraged to alternate between the role of tutor and tutee. Clearly, this issue manifested itself during the SP encounter. While all students obtained some practice, support, and ideas for working within a peer coaching model during the tutorial sessions of the Health and Social Behaviour in Physiotherapy Unit, further practice was needed to make the working relationships in a RPC model run a bit smoother.

The students' verbalisations also indicated that they were very good at recognising and affirming appropriate action and recognising and critiquing inappropriate action. This is a powerful source of potential learning for the students as it provides opportunities for confirming and developmental feedback. What was interesting, however, was that the peer coach rarely seemed to take the opportunity to provide confirming or developmental coaching feedback to their peer - particularly with the patient present. The peer coach would often verbalise that they had thought about raising the issue at the time but chose not to because they did not want to interfere with the rapport between their peer and the patient.

It is possible, however, that the students provided each other with their summative feedback during the discussion session or after the testing session was over. Unfortunately, this was not monitored in this study but would be useful for future studies investigating this same issue. Nonetheless, the delivery of useful coaching feedback was not usually delivered at the time when it would have been most useful - during the actual patient encounter. Clearly, students need further practice in developing this skill so that it can be provided in a more timely manner without jeopardising the client - therapist relationship.

The presence of peer support was another source of frequent verbalisation among the students in the RPC group. This peer support likely helped in keeping learning anxiety levels down to a manageable level in the RPC group. Johnson, Johnson, and Smith (1998), for example, have noted in their review of the literature that subjects in PAL initiatives generally report higher levels of support than subjects in individual models of learning with positive effect sizes in the order of 0.53.

8.2 Limitations

8.2.1 The Simulated Patient and the Simulation

The simulated patient (SP) was an effective method for introducing experimental control. This enabled a rigorous investigation of two different clinical education learning models. What was noteworthy about the outcomes of this particular simulation was that it was an extended patient case. As noted in the literature, most patient simulations occur within the context of an OSCE and usually have a duration of five to ten minutes. Maintaining accuracy and reliability of the SP's performance during these short encounters is much easier than is the case in a longer simulation. The fact that this SP was able to maintain a high degree of accuracy and reliability during each performance provides support for the use of prolonged simulations in research.

Both physiotherapists and physiotherapy students found the SP case to be very realistic, as evidenced by the positive ratings in the pilot study. The patient's history, physical signs and symptoms, and psycho-social attributes were all reproduced convincingly. Barrows (1993) has stated that SPs are quite capable of reproducing these features and the findings of this study support his conclusions. The high fidelity

of the patient simulation, therefore, strengthen the conclusions that have been made from this study as it can be argued that students appeared to perform in a genuine manner. This high fidelity supports the criterion validity of the SP as the simulation appeared to be able to elicit genuine performance from both physiotherapists and undergraduate physiotherapy students.

The accuracy and reliability of the SP's ratings also support the large body of evidence in the literature which states that well trained lay people can effectively evaluate candidate performance. In this study, the SP was able to rate the students' history and physical examination skills in a manner that closely paralleled the ratings of the investigator: an expert. The inter-rater reliability coefficients that were reported in this study fall within the upper ranges of what has been reported in the literature. For example, the SP in this study obtained an overall inter-rater reliability co-efficient of 0.84. Reports in the literature describe inter-rater reliability co-efficients in the range of 0.26 - 0.95.

It is important to note that a large proportion of the disagreements in the inter-rater reliability co-efficient stemmed from difficulties in the assessment of palpation. Quite often the investigator could not see exactly where a subject was palpating the patient. For example, the subject may turn their back on the investigator thus blocking his view of the technique momentarily. Jain et al. (1998), for example, found that small joint assessment could not be viewed as readily as other procedures in an OSCE thus reducing inter-rater reliability scores for the station. The intensity of the palpation pressure could also only be assessed by the simulated patient. Without this tactile information, and the occasional blockage in visual field, it was difficult to obtain concurrence in ratings between the SP and the investigator 100 per cent of the time.

There were also more discrepancies in the SP's ratings of the students' interviewing and communication skills in comparison to the ratings of the investigator. This discrepancy, however, is in keeping with reported trends in the literature. Individual bias often creeps into the evaluation of interpersonal communication skills because it is such a personal event. Nonetheless, the inter-rater reliability co-efficient that the SP obtained for his ratings of communication skill was 0.69. This is certainly in the

upper ranges of what has been reported in the literature. Ratings of communication by SPs fall within the range of 0.05 to 0.77.

The SP was also consistent as a rater over time. The intra-rater reliability co-efficient of the SP was 0.72 overall. The decrease in this value from the original inter-rater reliability score is due to observing the second situation on videotape. This is a contextually different experience from the original event as the SP can dedicate his focus solely on scoring. This decrease in intra-rater reliability from original inter-rater reliability scores has been reported in the literature. Tamblyn, Klass, Schnabl, and Kopelow (1991), for example, have stated that 12 per cent of findings in one study could not be assessed because of filming impediments. Within this particular study on RPC, similar difficulties with the use of videotape were noted. Fifteen per cent of the decrease in the SP's intra-rater reliability score could be attributed to filming impediments in this study.

In conclusion, the SP and the simulation itself was successful. The results of this study suggest that a well designed SP case, appropriately structured checklists and adequate SP training can lead to an effective simulation. This provides great benefits for educators and researchers needing a valid and reliable mechanism to effectively investigate differences in candidate performance.

However, this study employed an overt SP and was an experiment. Students in both groups verbalised incidents where they recognised that the encounter was artificial or simulated (Figure 12.13). Surprisingly though, these comments were not of the type which labeled the simulation as fake or unrealistic. Instead, students would be thrown off track somewhat because they could not believe that a SP could actually demonstrate pathology. When the SP presented genuine pathophysiological features to the students, which made sense from the perspective of the student's biomedical knowledge, it was quite a surprise. Again, this attests to the high fidelity of the SP and provides support to the argument that students performed in a genuine manner during the encounter.

Nonetheless, it is difficult to conclude with certainty that the students behaved like they would in real life. Because students knew they were being observed and evaluated, they were likely demonstrating competence and not necessarily their true

performance. Hence, their performance in the clinical setting could be quite different. However, the problems of replicating the performance of a real patient within the context of an actual health care system is difficult and necessitates the use of simulations if experimental studies of clinical learning are to have merit.

The use of an extended SP case also poses some minor limitations. It has been noted in the literature that longer stations yield more measurement error and less reproducible scores (van der Vleuten & Swanson, 1990). However, they also offer more content validity and provide opportunities for collecting more information on candidate performance. Hence, extended SP cases are useful for measuring inquiry strategies, sequencing of questions and skills, data analysis and diagnostic skill (Ferrell, 1995; Rosebraugh et al., 1996). This SP was very accurate and reliable and while there may have been measurement errors between candidates, the simulation was well controlled.

8.2.2 First Visit Bias

It is unlikely that there was a significant amount of ‘first visit bias’ influencing the outcome scores across the candidates in the areas of history taking and physical examination and, indeed, if there was, it would have applied to both groups equally. Nonetheless, all students were asked after the SP encounter to note in writing if there was anything they had chosen not to perform during the study. One consistent trend appeared. Students said they would have checked the middle and lower trapezius muscles in a subsequent visit when the patient had less pain and better shoulder range of motion. This was clearly a first visit bias because these tests were on the physical examination checklist but were not carried out by most of the students because it seemed inappropriate at the time to perform them. This also illustrates one example of how standard setters can sometimes impose their requirement for thoroughness even though in actual practice it would be inappropriate to carry out these specific maneuvers.

If the students did note the need to evaluate the middle and lower trapezius muscles they received a credit for considering this action. This minimised the impact of first visit bias. Other than this particular aspect of the physical examination, however, there did not appear to be any other systemic instances of first visit bias. Students

were able to carry out all of their desired activities within the context of the case and the time that was allowed. Only one dyad and one individual student required the full 60 minutes for the SP encounter.

8.2.3 Gender Bias

It unlikely that there was a gender bias in the simulation. Male and female students were distributed equally across the IND and RPC groups. Any gender effect would have been washed out or minimized by this assignment. The case that was designed was also innocuous from a gender perspective. The fact that the client was a male further minimised any problems of having to disrobe in front of the students. Had the SP been a female this may have influenced the scores received by male students. Unfortunately, it is difficult to measure actual differences in scores for male and female students in this study. This is largely due to the problem of extracting separate male and female scores from the SP encounter in the RPC group as these scores were derived as a composite.

8.2.4 Grapevine Effect

In order to have confidence in the results of this study, one must be certain that the IND group were responsible for working through the SP encounter independently and without influence from other peers. Similarly, one must be certain that the RPC group worked through the SP encounter as a team. It is possible, however, that students were collaborating with one another across the IND and RPC groups throughout the course of the study, thus creating a grapevine effect. If this was the case, then the effects of working individually versus cooperatively would be minimised as the entire class would be, in effect, peer coaching one another. Fortunately, the implementation of systems to control the possibility of a grapevine effect seemed to work in this particular study. The controls that were used were: having students sign confidentiality statements; reminding them of plagiarism and student code policies; telling them that the SP's findings may change over the course of the examination so it was best not to talk among peers because it could be misleading; ensuring that all checklists were secure so the students did not know what specific items of their performance were being evaluated; warning them that they may make reasoning errors if they have pre-conceived ideas about the case; and

factoring in some intergroup competition. Students were also told that if they had information about the SP it would make the simulation less realistic and they would not get the full benefit of the learning experience.

There were no statistically significant increases in history and physical examination checklist scores or PEQ scores over the four weeks of the study. In fact, scores tended to show a slight decrease over the four weeks. This downward trend in scores, however, was slight, except for mean history checklist scores for students in weeks one and three. As such, this specific difference being so slight is not of any practical significance. The lack of a grapevine effect in this study is consistent with reports in the literature. Most studies that have investigated grapevine effects report that these effects do not seem to appear in OSCE formats. This is largely due to the fact that students do not usually know the specific checklist details which form the basis of the examination. Hence, it is difficult to inform other students appropriately of what actions they should take during the examination.

It would be naive, however, to assume that students did not talk among themselves to a small degree. For example, students may have shared their ideas about the SP's diagnosis with one another. This may have influenced students in the latter part of the study, resulting in a more narrow and focused history and physical examination. Lower data collection scores, as a consequence of this narrower approach, would have occurred as the overall encounter would not have been as thorough as the approach of those candidates in the earlier part of the study who had no information. Evidence for this change in scores as a result student discussion has been outlined in the literature (Swartz, Colliver, Cohen, & Barrows, 1994).

8.2.5 The 15 Minute Discussion Period

In retrospect, it would have been useful to record the 15 minute discussion period for the RPC dyads in the qualitative aspect of this study. This would have provided a much more enriched understanding of some of the meta-cognitive discussions between students. However, having their thoughts tape recorded may have also hindered the discussion between students. This could have had a negative effect on their post-encounter CR test scores. In the end, the investigator decided not to

interfere with this part of the CR process. Future studies, however, should attempt to investigate the structure and dynamics of this peer interchange.

8.2.6 Retrospective Recall

One aspect of the qualitative methods used in this study relates to the retrospective recall method employed in this investigation. It is unrealistic to assume that the recall data is a wholly accurate description of the students' thoughts during the SP encounter. The 15 minute reflection interval, peer discussion, the information in the PEQ, and the process of observing the videotape itself all would have some influence on the integrity of the verbal recall data. Again, the decision to use a retrospective recall approach with the stimulus of the original episode was one made by the investigator. This emanated from concerns that students would find it difficult talking aloud while engaged in the task, particularly given their concern about not disclosing things in front of the patient. Further, numerous other studies have used the retrospective recall method with success. Ensuring the recall immediately followed the SP encounter and testing also helped to lessen problems with data integrity.

The use of stimulated recall proved to be an effective means of gaining a deeper understanding of the students' clinical reasoning. It also provided opportunities to evaluate learning issues within and across the IND and RPC models. The use of stimulated recall appeared to be able to draw out verbalizations that reflected the students' thinking at the time of the SP encounter. The integrity of this data was facilitated by having the recall sessions immediately after the SP encounter and testing session. One must recognize, however, that the discussion period and PEQ may have influenced the content of subsequent recall information. At best, the recall data is likely a mixture of level one and level two verbalisations.

As noted in Chapter 4, level one verbalisations represent a subject's internal speech and are assumed to be direct verbalisations of what is stored in short term memory (Patel & Arocha, 1995). Level two verbalisations are descriptions of recently acquired information which have been influenced somewhat by cognitive processes such as scanning, filtering and inference (Henry, Le Breck, & Holzemer, 1989). The relationship between heeded and verbalised information is still quite high but the verbal records are likely to possess a mixture of information from both the short term

and long term memory (Ericsson & Simon, 1993; Yinger, 1986). This mixture imposes somewhat of a limitation on the validity of the verbal recall data in this study as it is not entirely representative of the thinking that was taking place during the actual SP encounter. However, it still represents the students' final CR of the case and offers insights into differences among students in the IND and RPC groups.

The preparatory training for the recall sessions was effective in minimising the tendency for students to critique their performance on videotape. Viewing a videotape of one's performance is quite different from the actual event and one can not help but be somewhat evaluative when observing one's performance. While this did occur on occasion, it was very obvious and the students could be corrected. If a critique did appear in the students' verbalisations and subsequent transcription, it was not coded as it was not representative of their thoughts during the SP encounter.

The verbal prompts that were used by the investigator (i.e., "what were you thinking?") also kept the students on track during the recall sessions and reinforced the need to recall only what they were thinking at the time. Ensuring that the students' stopped the videotape once they recalled a thought also ensured that the complete verbalization was captured.

Other factors that must be taken into consideration when analysing the integrity of the verbal recall data are social desirability and evaluation apprehension. Although the stimulated recall session was completely separate from the SP encounter and not used in any way to calculate a student's grade, a student may still have been reluctant to say certain things during the recall session because the investigator is a member of the teaching staff and the students were still going to be graded in the unit. In the case of the RPC group, students may have also moderated the content of their recall in order to not offend their peer given that the recall sessions for the RPC group were carried out with both student peers present.

The recall sessions in the RPC group produced thoughts that could be related not only to CR but to the coaching experience as well. These latter thoughts were obviously present because both parties were cognisant of their peers presence throughout the encounter. In contrast, students in the IND group recalled thoughts that were limited mostly to their clinical reasoning.

8.2.7 Singular Case

Only one clinical condition or case was evaluated in this study. This obviously limits the generalisability of this study to the field of orthopaedics. Further investigations using other musculoskeletal cases would be helpful to increase the generalisability of the findings of this study.

The following chapter outlines the conclusions drawn from the current study. In particular, how the outcomes of this study relate to the literature on peer assisted learning and clinical reasoning.

Chapter 9: Conclusions

9.1 Reciprocal Peer Coaching - Relationship to the Literature

The results of this study on reciprocal peer coaching (RPC) appears to align itself with the purported benefits of peer assisted learning (PAL) that have been reported in the literature, in particular, the growing body of evidence in support of PAL in the higher education sector. The use of a RPC strategy, along with a cooperative goal structure, led to greater levels of achievement in novice physiotherapy students when compared to similar students in an individualistic (IND) learning experience. These higher levels of achievement that manifested in the reciprocal peer coaching (RPC) group appeared in the cognitive, psychomotor and affective domains. The theoretical underpinnings of these outcomes are discussed further in this final chapter.

Johnson et al. (1981) have noted that PAL is a useful teaching strategy for developing skills in tasks that require concept attainment and motor performance. Further, in a major review and meta-analysis of 168 studies on cooperative, competitive, and individual learning, cooperative learning strategies resulted in a positive mean effect size of 0.53 over individual learning in the tertiary education sector (Johnson, Johnson, & Smith, 1998). The results of this study are consistent with the outcomes reported by Johnson and colleagues. In this study, for example, the impact of a RPC strategy resulted in an overall positive effect size of 0.82 for the total patient encounter and 0.69 for the post-encounter clinical reasoning questionnaire. These are strong measures of practical significance and support the importance of introducing more PAL initiatives into the clinical learning setting.

The findings of this study also support the evidence that has been derived from the teacher staff development literature. This literature purports that when theory, demonstration and practice with non-evaluative feedback are combined with coaching, significant gains in achievement are noted (Ackland, 1991; Flynn, Bedinghaus, Snyder, & Hekelman, 1994; Hekelman et al., 1994; Johnson & Johnson, 1987; Joyce & Showers, 1995; Kohler, McCullough Crilley, & Shearer, 1997; Martin & Double, 1998; Showers, 1984; Showers, 1985; Williamson & Russell, 1990; Wynn & Kromrey, 1999). Peer coaching under these conditions lead to: a greater

awareness of theory and practice; changes in attitudes towards oneself, others and academic content; development of new skills; and the transfer of these skills. Several of these outcomes were identified in this particular study and illustrate that the peer coaching methods espoused in teacher education can be applied to other professions with good effects.

9.1.1 Reward Systems to Promote Peer Coaching

One factor that appeared to have a positive influence on the outcome of this study relates to the design of the learning experience. This study used within group cooperation and inter-group competition as the reward system. Students in the RPC groups knew that their overall score was going to be influenced by the performance of their peer. As a result, it was in their best interests to cooperate as a team. Had this incentive not been put into place, students may have chosen not to cooperate with their peer as there would have been an incentive to do better on their own. This raises important implications for educators both in the clinical and academic setting. Merely assigning students to work together may not result in the educational outcomes one was hoping to achieve. The influence of cooperative and competitive goal structures must be taken into consideration when designing curriculum.

Students in the RPC group (and IND group) also knew that their overall outcome score would be posted (by student number) alongside all the other outcome scores of their peers. This was done to provide the students with some feedback on their own performance in relation to their peers. Hence, there was somewhat of a competitive incentive to do well in relation to others. Lastly, the outcome scores for the simulated patient (SP) encounter were worth 25 per cent of the unit grade. This score was substantial enough for them to put some energy into the exercise. Collectively, therefore, these reward structures, which have been described in the literature, appear to have been of importance for the outcomes of this study.

To determine whether these reward systems were critical for obtaining the outcomes found in this investigation requires further study. Additional control groups would be needed to see whether the absence of group incentive and intergroup competition schemes produce the same amount of academic achievement. Obviously the question of group reward schemes and competition is worthy of further study given that an

educational objective of clinical education is proficiency in teamwork. Hence, the issue of inter-group competition and group rewards to facilitate performance in health science education remains controversial, particularly at the adult level where more mature individuals may respond differently to these motivating variables.

9.1.2 The Value of Discussion and Peer Interaction on Metacognition

The measurable outputs of this study indicate that RPC can be used to enhance achievement and reasoning. However, this does not necessarily explain the underlying theoretical reasons for this positive outcome. Goldschmid and Goldschmid (1976) describe several benefits of peer tutoring. Two of these benefits fall within the socio-psychological and pedagogical domain. Students provided strong socio-psychological benefits to one another during the RPC experience. Qualitative reports of peer support, and the quantitative reductions in anxiety seen in the RPC group, support this theoretical socio-psychological perspective. The pedagogical effects of peer assisted learning were also substantiated with the students in the RPC group. These students achieved significantly higher levels of performance and problem solving. The pedagogical benefits largely stem from the heightened metacognition that occurred during the clinical learning experience.

The pedagogical benefits that eventuated can be understood more readily by examining the cognitive development theories of Piaget (1977), Sullivan (1953) and Vygotsky (1986). These theories provide a framework for understanding how critical cognitive conflict supports the heightened performance of students in the RPC group (Piaget, 1977; Sullivan, 1953; Vygotsky, 1978; Vygotsky, 1986). These theorists argue that peer interaction is seen to promote cognitive development by creating critical cognitive conflicts. If the learner, through deliberations with their peer, becomes aware of a contradiction in their knowledge base, the experience creates a disequilibrating effect. This instigates the learner to question his or her beliefs and to try out new ones. Hence, if a student is seen to be following a certain line of patient care inquiry, and the other student does not understand the rationale behind this inquiry, disequilibrium occurs. The uninformed student, as a consequence, will initiate strategies to resolve this disequilibrium, for example, asking their peer for an explanation. The presence of disequilibrium was evident in

this study during the patient encounter and in the verbal reports of the students during the stimulated recall session. Students often asked questions of one another or sought clarification from one another to ensure their reasoning was on track.

More contemporary researchers also support this developmental perspective on cognitive controversy (Johnson, 1981; Johnson & Johnson, 1978; Johnson, Johnson, & Smith, 1986; King, 1997; Slavin, 1987). As was evident in this study, the intellectual disagreements that developed during the history and physical examination, and the post-encounter discussion, led to conceptual conflicts which motivated the learners in the RPC group to seek out new information. Students were able to self-manage these knowledge controversies effectively which led to the higher achievement levels evident in this group.

Johnson (1981) sees this cognitive controversy theory being comprised of two parts. The first part involves the conflict itself when one discovers that your own ideas do not appear compatible with that of your peer. The second part involves what Johnson (1981) terms epistemic curiosity. Epistemic curiosity involves an active search for more information. Both of these features increase in their utility when the disagreement is larger and more frequent. To promote controversy, Johnson (1981) recommends heterogeneity in the group such as age, gender and ability level differences. Unfortunately, the low:low, high:low and high:high groupings in this study were too small and do not provide enough power to determine whether a specific pairing strategy lead to higher achievement outcomes. Interestingly, Riggio, Whatley, & Neale (1994) did not discover any differences in groupings based upon academic level. They argue that the heightened cognitive activity that occurs among the pairs when disequilibrium is evident washes out any potential differences in academic ability. This may perhaps explain why low:low groups, in some situations, did better than high:high and high:low combinations.

The management of critical cognitive conflict also appears to be more amenable between peers because they speak on levels which can be easily understand by one another (Damon, 1984; Foot & Howe, 1998; King, 1997). The informal communications between peers are also less threatening than the corrective advice from a supervisor. Hence, levels of critical cognitive conflict may be lessened in

situations between learner and supervisor because of the power differential that is present in the relationship. Students may not be prepared to challenge a superior if they experience disequilibrium.

Damon (1984) and Foot and Howe (1998) also provide an overview of the theories of Vygotsky (1978,1986) and his social interactionist view of cognitive development. Vygotsky reports that peers benefit from one another by internalising the cognitive processes implicit in their interactions and communications. This is not dissimilar to epistemic curiosity and describes the heightened verbal discourse that takes place between learners locked in academic disagreement. The peer dialogue that occurred within the RPC group emulates several of the critical features of rational thinking described by Vygotsky (1978, 1986), in particular: the verification of ideas; the planning of strategies; the symbolic representation of intellectual acts; and the generation of new solutions. By engaging in this dialogue with a more capable peer, Vygotsky (1978, 1986) states that learners are able to enter the ‘zone of proximal development’. What this means is that learners can tap into the knowledge frameworks of their peers and use this discussion to enhance their own understanding of the problem that faces them, in this case, a patient examination. As noted earlier by Johnson (1981) ability level differences may be used to promote more structured controversy in learning. One of the reasons that low achieving students in the RPC group did better than low achieving student in the IND group may be that the former were able to enter new areas of potential competence.

Educators can tap into this cognitive developmental learning strategy by creating planned controversy in their teaching and learning experiences. By creating situations where learners will find their ideas in conflict, the motivation to resolve this conflict will increase. The bigger the disagreement the bigger the need to pursue the solution. Because students were accountable for understanding this case, cognitive conflict was unsettling and students were motivated to align their understanding of the case.

The cognitive developmental perspective is also useful for explaining how biomedical knowledge and theory can be directed towards improving actual clinical practice. The challenge of having your biomedical knowledge come under scrutiny by another peer is a useful method for stimulating metacognition as it requires a

student to think about how they think and to consider how much they know and do not know. This process is critical for the development of professional self-evaluation as it is this skill which enables learners to strategically pursue life long learning opportunities. The metacognitive processing that emanates from peer centered dialogue facilitates the development of more integrative and accurate clinical knowledge. This can be illustrated by expanding upon Boshuizen and Schmidt's (1992) model of clinical expertise. In stage two of this model, biomedical knowledge becomes encapsulated into more clinically relevant formats. During this second stage, the novice begins to develop and fine tune his/her propositions, networks and schemata. This occurs by processing similarities and differences in biomedical knowledge and aligning them with current clinical experience. The enhanced learning and performance outcomes of the RPC experience heightened the depth and accuracy of the learner's propositions, networks and schemata. Undoubtedly, this would facilitate transfer to future clinical practice, particularly when the cases are similar in nature.

In conclusion, the RPC model gives learners the opportunity to gain powerful insights into their knowledge gaps and mistakes and provides them with the chance to explore other alternatives. This further deepens their learning and understanding of the case and facilitates the development of their clinical competence. Educators should be interested in the RPC model as it provides an easy system for fostering the development of clinical competence in the health sciences sector. Quite often, students are expected to make the leap from the class room to the clinic. More often than not, however, educators are disappointed when they see that this does not happen as readily as they would like. The RPC method should be viewed as another way of encouraging more expansive thinking, cognition, meta-cognition and knowledge generation during clinical practice.

9.1.3 Learning From Peers

What is evident from this research on peer assisted learning is that the RPC model has the ability to help each student enhance his or her learning potential in a clinical learning environment. Both low and high achieving students obtained benefits from the RPC model. It is likely that these benefits emerged because of the high degree of

mutuality and equality in the learning experience (Damon & Phelps, 1989). Equality describes the extent to which learners take direction from one another. The nature of the RPC experience required student dyads to work closely together to ensure success. Mutuality describes the extent to which the learners' discourse is extensive, intimate and connected. The debriefing sessions that followed the patient encounter required the students to engage in useful discourse if they were to heighten their success in the post-encounter clinical reasoning questionnaire. Further, students had to develop communication strategies to support them during the patient encounter.

Deutsch (1949) noted that the way social interdependence is structured determines how individuals will interact. This structure, in turn, influences the amount of learning that actually takes place (Johnson et al. 1998). Power differences between supervisor and student can negatively influence learning. Adult learners need to feel safe and unthreatened when learning. Hence, the negative influence of power, whether real or perceived, needs to be mitigated for maximal learning outcomes. Since peers are at an equal level, feedback and guidance from a peer can be less threatening. With this greater perceived sense of safety, learners can be more open and inquisitive with one another.

The learning benefits that emerged in the RPC group were also not haphazard. These benefits were more likely to eventuate given the clear guidelines, rules and procedures for the learning experience. The fact that there was structure to the experience is important. Research by Fantuzzo (1989), Fantuzzo et al. (1989) and Riggio et al. (1991, 1994) demonstrate that performance is enhanced in dyadic group structures when a formalised structure is superimposed upon the learning experience. This theoretical perspective is certainly supported by the results of this research although there was not a parallel unstructured experience to test out this hypothesis more rigorously.

The theories of Bandura (1971, 1997) and the peer modelling views of Schunk (1998) provide a useful theoretical framework for examining the influence of PAL on learner efficacy from a social learning perspective. There were certainly many opportunities for direct reinforcement to occur as learners could get immediate feedback on their actions from their peers. Vicarious reinforcement or peer modelling was also a very strong modifier as the coach could directly observe the actions of

their peer and moderate their own actions based upon the outcomes they observed. All of these reinforcements, of course, contribute to the learners' metacognitive learning framework as they provide opportunities for identifying knowledge gaps and deficiencies and for re-constructing knowledge and practice.

High achieving students in the RPC model also obtained numerous benefits. Although much of the performance differential between high achieving students in the IND and RPC group were not statistically significant, there were still moderate effect sizes in favour of the latter group. High achieving students in the RPC group, however, did obtain considerable benefit in the area of clinical reasoning. The development of patient management plans, in particular, was one area where the peer discussion made a big difference. Clearly, the additional time spent in peer collaboration produced benefits for the high achieving student. Hence, participation in this learning model should be encouraged for both high and low achieving students.

One concern that is raised by clinical educators as a negative aspect of RPC is the possibility that students may give each other the 'wrong' information. This did not appear to be a big problem for this particular group of students although misinformation may have occurred during the 15 minute discussion session, which was not monitored. However, the possibility of misinformation in the RPC model is likely no worse than the possibility of no information in the IND model. In this latter learning model, students are often left to their own devices and never put into a position of cognitive disequilibrium. For example, individual students may be making inappropriate conclusions or suppositions that go completely unchecked. With the RPC model, peers appeared to be able to correctly affirm or critique actions. They rarely misinterpreted or intervened in such a way that a negative outcome was the result.

It is even plausible that students may learn more from their peers than their actual clinical supervisors. O'Donnell et al. (1985) has argued that peers provide a much greater volume of feedback which is generally more immediate than that which comes from supervisors. Immediacy is a central tenet of useful feedback. As a result, the volume and immediacy of feedback may make up for the few instances

where incorrect information is provided. Further, with real or perceived power differentials between student and supervisor, students may not even share their thoughts with their evaluator because of fears of negative appraisal. The limited availability of supervisors in most health care settings also means less immediacy and lower volumes of feedback for students. Hence, the RPC model offers some very useful learning benefits for learners, particularly in health care settings where access to more expert supervisors may be limited.

One potential issue that could arise in learning from peers is the consumer-oriented perspective of a student. Students may object to being put into the ‘unpaid’ teacher role, especially if a student recognises that their partner is less talented and knowledgeable. Extreme differences in student ability may not be the best learning arrangement. Resentment may develop in the relationship, particularly if the student with lower levels of ability becomes intimidated. Fortunately, this current research suggests that students appear to receive benefits from the RPC experience, regardless of their ability differentials. This may stem from the fact that the physiotherapy students in this study were not extremely different in ability from one another. This is reflected in their spread of academic scores in their orthopaedic science subjects (see Figure 6.2). The cognitive rehearsal that both high and low achieving students must engage in, when coaching their peer appears to augment their understanding and management of the patient case and may actually override ability differences. Positive changes in anxiety levels and enhanced clinical reasoning appeared in students across all levels of the ability spectrum. These outcomes should be attractive to students who are concerned about their learning outcomes in the RPC model.

9.1.4 Training

A consistent trend in the literature with respect to PAL is the importance of training for tutors or coaches. As was noted in Chapter 6, all students participated in a two hour tutorial on how to work successfully within a RPC dyad, regardless of their assignment to the IND or RPC group. During this session and others in the HSBIP Unit, students were given opportunities to explore issues such as joint decision making, communication, feedback, conflict management, and the link between peer learning and professional practice. Further, most students in this study would have

experienced working both within a group and independently, as their second year and third year clinical visits use a combination of individual and group learning. These group models, however, are relatively unstructured and students are basically organised into groups and expected to work collaboratively. Nonetheless it would have provided the students with some insights into group work.

In this particular study, the six students in the RPC group who were interviewed after the patient encounter reported that they recognized the need for further practice, training, and support to develop their collaborative skills. Actually spending some time thinking through how the dyad would work appeared to be an influential feature of a successful coaching experience. Even then, for those who dedicated some time to this planning function, they still found it difficult to coach their peer in front of the patient.

This lack of ‘active’ coaching in front of the patient was an interesting phenomenon. Students, as evidenced from the qualitative data, often affirmed or critiqued actions or were seeking clarification themselves. These critiques, affirmations and desires for clarification, however, were generally not shared with their peer at the time. While this rich source of learning and coaching information was present, it did not get brought into play at the time when it would be most effective. As has been noted in the literature, there may have been social influences at play which prevented them from correcting or critically evaluating each others’ arguments, particularly in front of the patient (Ackland, 1991). Clearly, more training, maturity, confidence or just more time and experience in working with a peer coach may be needed on the part of the students for active coaching to occur more regularly.

This raises an interesting point. When should programs introduce peer assisted learning strategies in the curriculum. Clearly, if students are expected to work cooperatively in academic and clinical environments, then the skills needed for successful practice must be integrated throughout the curriculum. This is particularly important for physiotherapy as students must compete vigorously against one another to gain entry into the academic program. The competitive skills that proved to be successful in gaining admission to the program must be re-evaluated, and possibly unlearned to some degree, if more collaborative and cooperative learning practices are to occur in the program.

9.1.5 Competition

Competition did appear between students in this particular study and it is important to raise this point as it is a very real concern in a RPC experience. As has been stated in the literature, learners in peer assisted learning situations may withdraw to work independently, take control or become the lead problem solver (Wynn & Kromrey, 1999). From this, jealousy, resentment and defensiveness may occur which interferes with the equality and mutuality of the group learning experience. Competitive behaviours appeared in the qualitative comments of the high:high group (Figure 7.60). However, it did not necessarily lead to a breakdown in performance. What did seem to appear was an increase in the number of critiques about the other student's actions and a desire to be in control so that one could clarify their own ideas. Table 7.34 illustrates this feature as the high:high group has many more incidents of critiquing action and seeking clarification.

9.2 The Teaching of Clinical Reasoning

Even though this study was a quasi-experimental investigation of the effects of a reciprocal peer coaching strategy on undergraduate physiotherapy students' problem solving skills, the results should be viewed from the perspective of clinical education programming. By implementing RPC models in clinical practice, students can develop deeper insights into their own personal CR and learn about the process of clinical problem solving in a much more structured way. Further, the heightened cognition and metacognition that emanates from cognitive controversy should enable students to develop clinical knowledge structures that serve as more effective prototypes for future practice.

The whole basis of RPC sits well with modern day theories of experiential learning (Boud, 1993; Boud & Edwards, 1999). Prior and current experience is seen to influence all learning, as learners bring the totality of their life history to the learning setting. Learning in these authentic situations allow concepts to evolve because the new situation, and the negotiations and discussions that occur, recast the information into a more densely textured form (Graham, 1996). Given that clinical knowledge is highly idiosyncratic, based upon the clinical experiences that a practitioner experiences throughout their career, the idea of experience based learning is appropriate for understanding the teaching and learning of clinical reasoning.

Through the practice of critical self-evaluation and elaboration with peers, each new experience that a student encounters can be related to prior experience. New knowledge can be re-organised using a variety of cognitive and metacognitive strategies that take into account the learner's critical self-evaluation and the elaboration that has occurred with their peer.

This perspective on learning is supported by educators in the health sciences (Boshuizen & Schmidt, 1992; Carr et al., 1995). They see the integration of general and context specific knowledge emanating from reflection and discussion. By exploring the connections between these two knowledge domains, encapsulation of biomedical knowledge into more relevant and robust clinical forms can take place. Further, by transforming biomedical knowledge into more useful clinical formats, transfer to future encounters may be enhanced.

Transfer of training is an interesting concept that may be facilitated by socio-cognitive learning approaches (Bandura, 1971; Bandura, 1997). Joyce and Showers (1982) have noted that an adequate background in theory, demonstration and practice does not necessarily translate into appropriate performance in context. Working alongside a peer is one mechanism of promoting this transfer as it provides companionship and support, technical feedback and analysis of executive control functions. Schunk (1998) calls this 'peer modeling' and has demonstrated how this technique can be used to promote transfer of training of skills, attitudes, beliefs and behaviours.

The use of RPC strategies to heighten the development of CR expertise also fits well with the contemporary views of expertise proposed by Boshuizen & Schmidt (1992, 1995), Schmidt and Boshuizen (1993) and Schmidt et al., (1990). This four stage theory postulates that expert performance is not due to superior reasoning skills or superior knowledge of patho-physiology, but rather, it is based on cognitive representations in memory that describe the features of prototypical patients. Strategies which enable learners to develop more accurate prototypical representations of their patients will undoubtedly facilitate the progression towards competency and expertise. For example, in stage one of this theory, the declarative knowledge of students is transformed into rich causal networks explaining the causes

and consequences of disease in terms of general underlying patho-physiological processes. These are called propositional networks and represent how things, objects, events or concepts relate to one another. As learning progresses, these causal networks become increasingly complex. The development of these networks is dependent upon understanding these events and concepts thoroughly. Inappropriate suppositions lead to errors in reasoning. Hence, the deliberations that occur in RPC can be highly effective in the development of finely tuned propositional networks.

In stage two, these knowledge networks are compiled and become abridged networks. These causal models, which delineate signs and symptoms, begin to become subsumed under diagnostic labels with repeated exposure to clients. Knowledge that is pertinent to the specific case becomes activated and CR begins to become more efficient. Schmidt and Boshuizen (1993) have studied the reasoning of students at this level. The verbal protocols these students produce are very detailed and contain many biomedical propositions. Students are actively processing their basic science knowledge in conjunction with their emerging clinical knowledge. These propositions, however, are often inadequately linked to post hoc explanations of the diagnosis. This increase in reasoning error for clinicians at stage two is called the 'intermediate effect' and is representative of the restructuring taking place in the clinician's biomedical and clinical knowledge. Again, the critical cognitive conflict and elaboration that occur in RPC learning models can minimise errors of reasoning common at this stage of development. Linkages to post hoc explanations of the diagnosis can be enhanced and the overall transformation of biomedical to clinical knowledge facilitated.

The development and fine tuning of these knowledge networks lead to the development of illness scripts. Illness scripts not only encompasses the biomedical aspects of the case, but the prognostic and treatment components as well. The theoretical notion of illness scripts to guide clinical reasoning were originally put forward by Feltovich and Barrows (1984) and expanded upon further by Boshuizen and Schmidt (1995). The use of RPC models may arguably facilitate the development of richer more elaborate illness scripts, or at least the prototypes that lead to illness scripts. Given that the learners in the RPC model were more thorough and presented

with a deeper understanding of the case, as evidenced in their PEQ scores, their mental models of the case and future cases will be more highly developed. Hence, their baseline prototypes and rudimentary illness scripts are more likely to be highly developed than individual learners who experience a similar case in isolation.

This progression towards more finely tuned propositional networks and illness scripts is supported by research in the field. Work by Bordage and Lemieux (1986) provides a strong theoretical basis for employing peer assisted learning strategies to promote clinical reasoning. They studied the CR skills of students in a problem-based curriculum and found that the stronger students had more highly developed knowledge structures with stronger relational semantic networks. Weaker students relied on syntactical knowledge structures, which resulted in fewer and weaker mental operations during the CR process. Encouraging students to discuss cases amongst themselves, therefore, may actually enhance the reasoning of both parties in a RPC model because of the opportunity to review, revise, and add to their existing relational knowledge networks. Again, this is akin to the views expressed by Vygotsky (1978, 1986) and his description of how learners engaged in peer collaboration enter the zone of proximal development of their peers.

The learning situation that takes place in a RPC model also encourages more free associative thought among the students. This encourages them to explore alternative perspectives without having to worry about whether or not the other person agrees with them. This is often quite different from interactions between student and supervisor where the student is reluctant to challenge the views of their evaluator. For example, Showers (1985) advocates placing coaching into the hands of peers to minimise status and power differentials that interfere with learning. The RPC model, however, does not mean that the students work in isolation from their supervisor. They still require feedback from their supervisor. Independent evaluation and feedback from a more experienced practitioner is still valuable for the learner, particularly when the dyad is unable to solve a problem or are failing to remediate poor performance.

The teaching of clinical reasoning using experiential learning principles fits well with current views of constructivist learning and metacognition. Joyce and Weil (1996)

support the importance of metacognition in learning from a constructivist perspective. Constructivists believe that learners actively construct their own reality or knowledge using personal experience to structure their rules, concepts, hypotheses and associations (Biehler & Snowman, 1997). This personalised knowledge is influenced by the learner's experiences, beliefs, gender, age, ethnic background and individual biases (Biehler & Snowman, 1997; Graham, 1996). Joyce and Weil (1996) challenge the view that knowledge and skill development must be provided to students in neat packages. Instead, they see knowledge development as personal, and occurring within a social perspective through interactions with teachers and other learners. Hence, each learner's concepts and meanings will be unique, and the learner's task, therefore, is to seek meaning within their own unique frame of reference. Building knowledge, and checking it against the concepts of others, is seen by to be a major part of the process of education (Biehler & Snowman, 1997; Joyce & Weil, 1996).

This constructivist learning perspective needs to be factored into clinical education programs much more strategically so that students can more readily shape their biomedical knowledge into clinical knowledge. In traditional apprenticeship models, students often end up mirroring the approaches of their clinical supervisors. This mirroring does not provide the student with a deep understanding of the concepts inherent in their approach. Shifting the responsibility to students, and making them accountable for developing their own approach, will deepen the learner's sense of professional autonomy. Further, by engaging in this process with a supportive peer, the potential for constructing an idiosyncratic knowledge base, which will support that learner in the future will be heightened.

From a health sciences perspective, therefore, the use of RPC appears to be as good, or better, than the traditional individualistic method of learning for developing the CR and performance skills of novice physiotherapy students. It supports the concepts of deep learning espoused by Biggs (1987, 1988). At best, the incorporation of both models into clinical education programs should take place so that learning and autonomous practice can both be maximised. The RPC model also emulates professional practice in the workplace. The health care system values teamwork.

Modeling this practice as part of the clinical learning experience provides students with rich opportunities to develop generic skills for the workplace.

9.3 Future Research Implications

One aspect of this study which was not investigated was transfer of knowledge or training. Peer coaching should, theoretically, encourage transfer of training because it forces learners to practice their new skills more frequently and appropriately with their coach. When learners are faced with a similar context or situation, this increased level of practice should be carried over to the new situation. This transfer of training is particularly useful in the health sciences. Procedural learning, or putting knowledge and skills into practice, relies heavily on receiving feedback about performance (Johnson & Johnson, 1987). By using non-evaluative feedback from peers in coaching situations, the learner can modify his/her actions until errors are eliminated. It would be interesting to expand upon this study to include a second patient with shoulder pathology and to have all students examine this patient independently. However, whether it would be ascertained that students from the RPC group would continue to outperform their peers from the IND group is difficult because of the problems of case specificity. Norman, Feightner, and Tugwell (1983), for example, found very little correlation in the performance scores of students for two identical stations that were utilised during an objectively structured clinical examination.

Further inquiry into how metacognition enhances the clinical performance and reasoning of students in a RPC model would also be beneficial. In this study, the verbal discussion among paired students that followed the patient encounter was not monitored. It would be extremely useful to evaluate the nature and quality of this discussion to map out those metacognitive strategies that students knowingly or unknowingly apply when engaged in clinical problem solving. For example, when discussing the patient case, are the students being self-regulatory or just task-oriented?

Clearly, further studies evaluating the benefits of RPC in other physiotherapy specialties are needed to determine the generalisability of these outcomes. Reciprocal

peer coaching appears to be as good, if not better, than the more traditional approach of individualistic learning that is frequently used in undergraduate physiotherapy clinical education. Higher levels of student achievement, reductions in learner anxiety, and deeper learning are all potential advantages of the RPC model. The use of a simulated patient to study learning issues in clinical education has also proved to be a highly effective strategy for conducting research in this field. Investigators in clinical education should take advantage of this technology to explore the many questions about clinical reasoning, clinical education, and clinical competence, that still need to be addressed in health professional education.

References

-
- Ackland, R. (1991). A review of the peer coaching literature. *Journal of Staff Development, 12*(1), 22-26.
- Adams, J., & Hamblen, D. (1990). *Outline of Orthopaedics*. (11th ed.). Edinburgh: Churchill Livingstone.
- AERA, APA, & NCME. (1985). *Standards for educational and psychological testing*. Washington D.C.: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.
- Ainsworth, M., Rogers, L., Markus, J., Dorsey, N., Blackwell, T., & Petrusa, E. (1991). Standardized patient encounters - a method for teaching and evaluation. *Journal of the American Medical Association, 266*(10), 1390-1396.
- Albanese, M., & Mitchell, S. (1993). Problem based learning: a review of literature on its outcomes and implementation issues. *Academic Medicine, 68*(1), 52-81.
- American Association of Medical Colleges. (1998). Emerging Trends in the Use of Standardized Patients. *Contemporary Issues in Medical Education, 1*(7).
- Ananthkrishnan, N. (1993). Objective structured clinical / practical examination (OSCE/OSPE). *Journal of Postgraduate Medicine, 39*(2), 82-84.
- Anderson, J., Reder, L., & Simon, H. (1996). Situated learning and education. *Educational Researcher, 25*(4), 5-11.
- Antil, L., Jenkins, J., & Wayne, S. (1998). Cooperative learning: prevalence, conceptualisations and the relations between research and practice. *American Education Research Journal, 35*(3), 419-454.
- APA. (1974). *Standards for educational and psychological tests*. Washington D.C: American Psychological Association.
- Arocha, J., Patel, V., & Patel, Y. (1993). Hypothesis generation and the coordination of theory and evidence in novice diagnostic reasoning. *Medical Decision Making, 13*(3), 198-211.
- Aviram, M., Ophir, R., Raviv, D., & Shiloah, M. (1998). Experiential learning of clinical skills by beginning nursing students: "coaching" project by fourth-year student interns. *Journal of Nursing Education, 37*(5), 228-231.
- Avis, W. (Ed.). (1980). *Funk & Wagnalls Standard College Dictionary* (Canadian Edition ed.). Toronto: Fitzhenry and Whiteside Limited.
- Balla, J., & Boyle, P. (1994). Assessment of student performance: a framework for improving practice. *Assessment and Evaluation in Higher Education, 19*(1), 17-28.
- Bandura, A. (1971). *Social Learning Theory*. New York: General Learning Press.
- Bandura, A. (1997). *Self Efficacy: The Exercise of Control*. New York: WH Freeman.

- Barker-Schwartz, K. (1991). Clinical reasoning and new ideas on intelligence: implications for teaching and learning. *American Journal of Occupational Therapy*, 45(11), 1033-1037.
- Barrows, H. (1971). *Simulated Patients (programmed patients)*. Springfield, Illinois: Charles C. Thomas.
- Barrows, H. (1984). *Newer approaches to the assessment of clinical performance*. Springfield, Illinois: Southern Illinois University.
- Barrows, H. (1987). *Simulated (standardised) patients and other human simulations*. Chapel Hill, North Carolina: Health Sciences Consortium.
- Barrows, H. (1990). Inquiry: the pedagogical importance of a skill central to clinical practice. *Medical Education*, 24, 3-5.
- Barrows, H. (1993). An Overview of the Uses of Standardized Patients for Teaching and Evaluating Clinical Skills. *Academic Medicine*, 68(6), 443-451.
- Barrows, H., & Abrahamson, S. (1964). The programmed patient: a technique for appraising student performance in clinical neurology. *Journal of medical education*, 39(August), 802-805.
- Barrows, H., & Bennett, K. (1972). The diagnostic (problem solving) skill of the neurologist. *Archives of Neurology*, 26, 273-277.
- Barrows, H., & Feltovich, P. (1987). The clinical reasoning process. *Medical Education*, 21, 86-91.
- Barrows, H., & Tamblyn, R. (1980). *Problem Based Learning: An Approach to Medical Education*. New York: Springer Publishing Company.
- Barrows, H., Norman, G., Neufeld, V., & Feightner, J. (1982). The clinical reasoning of randomly selected physicians in general medical practice. *Clinical and Investigative Medicine*, 5(1), 49-55.
- Barrows, H., Williams, R., & Moy, R. (1987). A comprehensive performance-based assessment of fourth-year students' clinical skills. *Journal of Medical Education*, 62(Oct.), 805-809.
- Battles, J., Carpenter, J., McIntire, D., & Wagner, J. (1994). *Stability of scores on an OSCE over time*. Paper presented at the Sixth Ottawa Conference on Medical Education, Toronto.
- Beeken, J. (1991). Cooperative learning: planning for success. *Journal of Ophthalmic Nursing and Technology*, 10(2), 66-68.
- Benner, P. (1982). From novice to expert. *American Journal of Nursing*, March, 402-407.
- Benner, P. (1984). *From Novice to Expert: Excellence and Power in Clinical Nursing Practice*. London: Addison-Wesley.
- Benner, P., & Tanner, C. (1987). Clinical judgement: how expert nurses use intuition. *American Journal of Nursing*, 87, 23-31.

- Benner, P., Tanner, C., & Chesla, C. (1992). From beginner to expert: gaining a differentiated clinical world in critical care nursing. *Advances in Nursing Science*, 14(3), 13-28.
- Berkson, L. (1993). Problem-based learning: have the expectations been met. *Academic medicine*, 68(10), S79-S88.
- Biehler, R., & Snowman, J. (1997). *Psychology Applied to Teaching*. (8th ed.). Boston: Houghton Mifflin Company.
- Biggs, J. (1987). *Student Approaches to Learning and Studying*. Melbourne: Australian Council for Educational Research.
- Biggs, J. (1988). The role of metacognition in enhancing learning. *Australian Journal of Education*, 32(2), 127-138.
- Biggs, J., & Rihn, B. (1984). The effects of intervention on deep and surface approaches to learning. In J. Kirby (Ed.), *Cognitive Strategies and Educational Performance* (pp. 279-293). London: Academic Press.
- Blackwell, T., & Callaway, M. (1992). *The influence of sub-specialty ward experience on OSCE performance*. Paper presented at the Approaches to the Assessment of Clinical Competence, Dundee, Scotland.
- Bloom, B. (1953). Thought processes in lectures and discussions. *Journal of General Education*, 7, 160-169.
- Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of Educational Objectives*. London: Longman Group Ltd.
- Bordage, G., & Lemieux, M. (1986). *Some cognitive characteristics of medical students with and without diagnostic reasoning difficulties*. Paper presented at the 25th Annual Conference of Research in Medical Education, New Orleans.
- Bordage, G., & Lemieux, M. (1991). Semantic structures and diagnostic thinking of experts and novices. *Academic Medicine*, 66(9), S709-S772.
- Bordage, G., & Page, G. (1987). An alternate approach to PMPs: The key features concept. In I. Hart & R. Harden (Eds.), *Further Developments in Clinical Competence* (pp. 57-75). Montreal: Can-Heal Publications.
- Bordage, G., & Zacks, R. (1984). The structure of medical knowledge in the memories of medical students and general physicians: categories and prototypes. *Medical Education*, 18, 406-416.
- Borg, W., & Gall, M. (1983). *Educational Research: An Introduction*. (4th ed.). New York: Longman Inc.
- Boshuizen, H., & Schmidt, H. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science*, 16, 153-184.
- Boshuizen, H., & Schmidt, H. (1995). The development of clinical reasoning expertise. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 24-32). Oxford, UK: Butterworth Heinemann Ltd.

- Boud, D. (1988). How to help students learn from experience. In K. Cox & C. Ewan (Eds.), *The Medical Teacher* (2nd ed., pp. 68-73). London: Churchill Livingstone.
- Boud, D. (1993). Experience as the base for learning. *Higher Education Research and Development*, 12(1), 33-44.
- Boud, D., & Edwards, H. (1999). Learning for practice: promoting learning in clinical and community settings. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners: Challenges for Health Professional Education* (pp. 173-179). Oxford: Butterworth-Heinemann.
- Boulet, J., Friedman Ben-David, M., Amitai, Z., Burdick, W., Curtis, M., Peitzman, S., & Gary, N. (1998). Using standardized patients to assess the interpersonal skills of physicians. *Academic Medicine*, 73(10), S94-S96.
- Bowman, M., Russell, N., Boekeloo, B., Rafi, I., & Rabin, D. (1992). The effect of educational preparation on physician performance with a sexually transmitted disease-simulated patient. *Arch Intern Med*, 152(September), 1823-1828.
- Brook, W. (1997). *Standardised Patient Protocols: Impingement Syndrome & Supraspinatus Tendinitis* : University of Alberta.
- Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32-42.
- Burri, A., McCaughan, K., & Barrows, H. (1976). *The feasibility of using the simulated patient as a means to evaluate clinical competence of practicing physicians in a community (a pilot project)*. Paper presented at the Fifteenth Conference in Medical Education, Washington, D.C.
- Caillet, R. (1981). *Shoulder Pain*. (2nd ed.). Philadelphia: F.A. Davis.
- Calhoun, J., Woolliscroft, J., & Ten Haken, J. (1987). Internal medicine house officers' performance as assessed by experts and standardized patients. *Journal of Medical Education*, 62(September), 754-760.
- Candy, P., & Worrall-Carter, L. (1999). Educating health science students for life long learning. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners: Challenges for Health Professional Education* (pp. 160-165). Oxford: Butterworth-Heinemann.
- Carnevali, D. (1995). Self-monitoring of clinical reasoning behaviours: promoting professional growth. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 179-190). Oxford, UK: Butterworth Heinemann Ltd.
- Carpenter, C. (1996). The evolving culture of physiotherapy. *Physiotherapy Canada*, 48(1), 11-15.

- Carr, J., Jones, M., & Higgs, J. (1995). Teaching towards clinical reasoning expertise in physiotherapy practice. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 235-245). Oxford, UK: Butterworth Heinemann Ltd.
- Carroll, M. (1996). Peer tutoring: can medical students teach biochemistry? *Biochemical Education*, 24(1), 13-15.
- Case, S., & Swanson, D. (1993). Extended-matching items: A practical alternative to free-response questions. *Teaching and Learning in Medicine*, 5(2), 107-115.
- Cason, C., Cason, G., & Bartnik, D. (1977). Peer instruction in professional nurse education: a qualitative case study. *Journal of Nursing Education*, 16(7), 10-22.
- Champion, P. (1989). *An evaluation of the use of simulated patients in health science education*. Unpublished Master's, La Trobe University.
- Chapman, E. (1998). Key considerations in the design and implementation of effective peer-assisted learning programs. In K. Topping & S. Ehly (Eds.), *Peer Assisted Learning 1998* (pp. 67-84). London: Lawrence Erlbaum and Associates.
- Chapparo, C., & Ranka, J. (1995). Clinical reasoning in occupational therapy. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 88-102). Oxford, UK: Butterworth Heinemann Ltd.
- Chase, W., & Simon, H. (1973a). The mind's eye in chess. In W. Chase (Ed.), *Visual information processing* (pp. 215-281). New York: Academic Press.
- Chase, W., & Simon, H. (1973b). Perception in chess. *Cognitive Psychology*, 4, 55-81.
- Chi, M., Feltovich, P., & Glaser, R. (1981). Representation of physics knowledge by experts and novices. *Cognitive Science*, 5, 121-152.
- Christensen, N. (1993). *Clinical pattern recognition in physiotherapists - a pilot study investigating the effect of different levels of experience*. Unpublished Master of Applied Science, University of South Australia.
- Cinelli, B., Wolford-Symons, C., Bechtel, L., & Rose-Colley, M. (1994). Applying cooperative learning in health education practice. *Journal of School Health*, 64(3), 99-102.
- Claessen, H., & Boshuizen, H. (1985). Recall of medical information by students and doctors. *Medical Education*, 19, 61-67.
- Coakes, S., & Steed, L. (1997). *SPSS: Analysis Without Anguish Version 6.1 for IBM and Macintosh users*. Brisbane: John Wiley Ltd.
- Coggan, P., Knight, P., & Davis, P. (1980). Evaluating students in family medicine using simulated patients. *Journal of Family Medicine*, 10(2), 259-265.

- Cohen, D., Colliver, J., Marcy, M., Fried, E., & Swartz, M. (1996). Psychometric properties of a standardized-patient checklist and rating scale form used to assess interpersonal and communication skills. *Academic medicine*, 71(1 Supplement), S87-S89.
- Cohen, D., Colliver, J., Robbs, R., & Swartz, M. (1997). A large-scale study of the reliabilities of checklist scores and ratings of components of interpersonal and communication skills evaluated on a SP examination. In A. Scherpbier, C. van der Vleuten, J. Rethans, & A. van der Steeg (Eds.), *Advances in Medical Education* (pp. 424-426). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Cohen, D., Gromoff, B., & Swartz, M. (1992). *Standardized patients - the advantages of the professional actor*. Paper presented at the Approaches to the Assessment of Clinical Competence, Dundee, Scotland.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- Cohen, J. (1969). *Statistical Power Analysis for the Behavioural Sciences*. New York: Academic Press.
- Cohen, R., & Sampson, J. (1999). Working together: students learning collaboratively. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners: Challenges for Health Profession Education* (pp. 204-211). Oxford: Butterworth-Heinemann.
- Cohn, E. (1989). Fieldwork education: shaping a foundation for clinical reasoning. *American Journal of Occupational Therapy*, 43(4), 240-244.
- Colliver, J., & Williams, R. (1993). Technical Issues: Test Application. *Academic Medicine*, 68(6), 454-460.
- Colliver, J., Barrows, H., Vu, N., Verhulst, S., Mast, T., & Travis, T. (1991). Test security in examinations that use standardised patient cases at one medical school. *Academic Medicine*, 66(5), 279-282.
- Colliver, J., Marcy, M., Travis, T., & Robbs, R. (1991b). The interaction of student gender and standardised-patient gender on a performance-based examination of clinical competence. *Academic Medicine*, 66(9), S31-S33.
- Colliver, J., Robbs, R., & Vu, N. (1991c). Effect of using two or more standardised patients to simulate the same case on case means and case failure rates. *Academic Medicine*, 66(10), 616-618.
- Colliver, J., Vu, N. V., Marcy, M., Travis, T., & Robbs, R. (1993). Effects of examinee gender, standardised-patient gender, and their interaction on standardised patients' ratings of examinees' interpersonal and communication skills. *Academic Medicine*, 68(2), 153-157.
- Connell, K., Sinacore, J., Schmid, F., Chang, R., & Perlman, S. (1993). Assessment of clinical competence of medical students by using standardised patients with musculoskeletal problems. *Arthritis and Rheumatism*, 36(3), 394-400.
- Cook, L. (1991). Cooperative learning: a successful college teaching strategy. *Innovative Higher Education*, 16(1), 27-38.

- Corcoran, S. (1986a). Expert and novice nurses' use of knowledge to plan for pain control. *American Journal of Hospice Care*, 3(Nov/Dec), 37-41.
- Corcoran, S. (1986b). Planning by expert and novice nurses in cases of varying complexity. *Research in Nursing and Health*, 9, 155-162.
- Corcoran, S., Narayan, S., & Moreland, H. (1988). "Thinking aloud" as a strategy to improve clinical decision making. *Heart and Lung*, 17(3), 463-468.
- Corcoran-Perry, S., & Narayan, M. (1995). Teaching clinical reasoning to nurses in clinical education. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 263-265). Oxford, UK: Butterworth Heinemann Ltd.
- Corsini, R. (Ed.). (1984). *Encyclopedia of Psychology*. (Vol. 2). New York: John Wiley & Sons.
- Costello, J. (1989). Learning from each other: peer teaching and learning in student nurse training. *Nurse Education Today*, 9, 203-206.
- Cotton, J., & Cook, M. (1982). Meta-analyses and the effects of various systems: Some different conclusions from Johnson et al. *Psychological Bulletin*, 92, 176-183.
- Coughlin, L., & Patel, V. (1987). Processing of critical information by physicians and medical students. *Journal of Medical Education*, 62, 818-828.
- Crabtree, M., & Lyons, M. (1997). Focal points and relationships: a study of clinical reasoning. *British Journal of Occupational Therapy*, 60(2), 57-64.
- Croen, L., & Moroff, S. (1994). Pilot-testing a holistic approach to scoring performance on standardised-patient examinations. *Academic Medicine*, 69(4), 310-312.
- Currier, D. (1990). *Elements of Research in Physical Therapy*. (3rd ed.). Baltimore: Williams and Wilkins.
- Damon, W. (1984). Peer education: the untapped potential. *Journal of Applied Developmental Psychology*, 5, 331-343.
- Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research*, 13, 9-19.
- Davis, C. (1998). *Patient practitioner interaction: an experiential manual for developing the art of health care*. (3rd ed.). New Jersey: Slack, Inc.
- Dawson-Saunders, B., Verhulst, S., Marcy, M., & Steward, D. (1987). *Variability in standardised patients and its effect on student performance*. Paper presented at the Further developments in assessing clinical competence, 2nd Conference, Montreal, Canada.
- Day, R., Hewson, M., Kindy, P., & Van Kirk, J. (1993). Evaluation of Resident Performance in an Outpatient Internal Medicine Clinic Using Standardized Patients. *Journal General Internal Medicine*, 8, 193-198.
- De Champlain, A., Margolis, M., King, A., & Klass, D. (1997). Standardized patients' accuracy in recording examinees' behaviours using checklists. *Academic Medicine*, 72(10), S85-S87.

- de Groot, A. (1965). *Thought and choice in chess*. The Hague: Monton.
- DeClute, J., & Ladyshevsky, R. (1993). Enhancing clinical competence using a collaborative clinical education model. *Physical Therapy*, 73(10), 683-689.
- DeDea, L. (1996). *The process, design, and implementation of an alternative, collaborative approach to clinical education using the 3:1 supervisory model*. Paper presented at the 12th International Congress of the World Confederation for Physical Therapy, Washington, D.C.
- Dekkar, J., van Baar, M., Curfs, E., & Kerssens, J. (1993). Diagnosis and treatment in physical therapy: an investigation of their relationship. *Physical Therapy*, 73(9), 568-577.
- Delitto, A., & Snyder-Mackler, L. (1995). The Diagnostic Process: Examples in Orthopedic Physical Therapy. *Physical Therapy*, 75(3), 203-211.
- Denzin, N. (1983). Interpretive Interactionism. In G. Morgan (Ed.), *Beyond Method: Strategies for Social Research* (pp. 129-146). Beverly Hills: Sage.
- Dettmann, M., & Linder, M. (1988). Significance of prior experience on students' clinical performance. *Journal of Physical Therapy Education*, 2(1), 7-9.
- Deutsch, M. (1949). A theory of cooperation and competition. *Human Relations*, 2(2), 129-152.
- Edwards, H., Franke, M., & McGuinness, B. (1995). Using simulated patients to teach clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 269-277). Oxford, U.K.: Butterworth-Heinemann Ltd.
- Edwards, I., Jones, M., Carr, J., & Jensen, G. (1998). *Clinical reasoning in three different fields of physiotherapy; a qualitative study*. Unpublished Masters, University of South Australia.
- Edwards, M., & Martin, A. (1989). The objective structured clinical examination as a method of occupational therapy student evaluation. *Canadian Journal of Occupational Therapy*, 56(3), 128-131.
- Eli, I. (1996). Reducing confirmation bias in clinical decision making. *Journal of Dental Education*, 60(10), 831-835.
- Elliot, D., & Hickam, D. (1987). Evaluation of physical examination skills: Reliability of faculty observers and patient instructors. *JAMA*, 258(23), 3405-3408.
- Elstein, A. (1995). Clinical Reasoning in Medicine. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 49-59). Oxford, UK: Butterworth Heinemann Ltd.
- Elstein, A., Kagan, N., & Shulman, L. (1972). Methods and theory in the study of medical inquiry. *Journal of Medical Education*, 47, 85-92.
- Elstein, A., Loupe, M., & Erdmann, J. (1971). An experimental study of diagnostic thinking. *Journal of Structural Learning*, 2(4), 45-53.
- Elstein, A., Shulman, L., & Sprafka, S. (1978). *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, Massachusetts: Harvard University Press.

- Elstein, A., Shulman, L., & Sprafka, S. (1990). Medical Problem Solving: A Ten Year Retrospective. *Evaluation and the Health Professions, 13*(1), 5-36.
- Embrey, D., Guthrie, M., White, O., & Dietz, J. (1996). Clinical decision making by experienced and inexperienced pediatric physical therapists for children with diplegic cerebral palsy. *Physical Therapy, 76*(1), 20-33.
- Ericsson, K., & Simon, H. (1984). *Protocol Analysis: Verbal Reports as Data*. (First ed.). Cambridge, Massachusetts: Harvard University Press.
- Ericsson, K., & Simon, H. (1993). *Protocol Analysis: Verbal Reports as Data*. (Revised Edition ed.). Cambridge, Massachusetts: The MIT Press.
- Fabiny, A., McArdle, P., Perls, T., Inui, T., & Sheehan, M. (1998). The geriatric objective structured clinical exercise: a teaching tool in a geriatrics curriculum. *Gerontology and Geriatrics Education, 18*(4), 63-70.
- Fantuzzo, J. (1989). Reciprocal peer tutoring: developing and testing effective peer collaborations for elementary school students. In K. Topping & S. Ehly (Eds.), *Peer Assisted Learning* (pp. 121-144). London: Lawrence Erlbaum and Associates.
- Fantuzzo, J., Riggio, R., Connelly, S., & Dimeff, L. (1989). Effects of reciprocal peer tutoring on academic achievement and psychological adjustment: a component analysis. *Journal of Educational Psychology, 81*(2), 173-177.
- Feltovich, P., & Barrows, H. (1984). Issues of generality in medical problem-solving. In H. Schmidt & M. de Volder (Eds.), *Tutorials in Problem-Based Learning*. Assen, The Netherlands: Van Gorcum.
- Ferguson, K., & Edwards, H. (1999). Providing clinical education: the relationship between health and education. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners: Challenges for Health Professional Education* (pp. 52-58). Oxford: Butterworth-Heinemann.
- Ferrell, B. (1995). Clinical performance assessment using standardized patients: A primer. *Family Medicine, 27*(1), 10-15.
- Flavell, J. (1979). Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906-911.
- Fleming, M. (1991a). Clinical reasoning in medicine compared with clinical reasoning in occupational therapy. *American Journal of Occupational Therapy, 45*, 988-996.
- Fleming, M. (1991b). The therapist with the three track mind. *American Journal of Occupational Therapy, 45*, 1007-1014.
- Flynn, S., Bedinghaus, J., Snyder, C., & Hekelman, F. (1994). Peer coaching in clinical teaching: a case report. *Family Medicine, 26*(9), 569-570.
- Foley, M., Nespoli, G., & Conde, E. (1997). Using standardized patients and standardized physicians to improve patient-care quality: results of a pilot study. *The Journal of Continuing Education in Nursing, 28*(5), 198-204.

- Fonteyn, M. (1995). Clinical reasoning in nursing. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 60-71). Oxford, UK: Butterworth Heinemann Ltd.
- Fonteyn, M., & Fisher, A. (1995). Use of Think Aloud Method to Study Nurses' Reasoning and Decision Making in Clinical Practice Settings. *Journal of Neuroscience Nursing*, 27(2), 124-128.
- Fonteyn, M., Kuipers, B., & Grobe, S. (1993). A Description of Think Aloud Method and Protocol Analysis. *Qualitative Health Research*, 3(4), 430-441.
- Foot, H., & Howe, C. (1998). The psychoeducational basis of peer assisted learning. In K. Topping & S. Ehly (Eds.), *Peer Assisted Learning* (pp. 27-43). London: Lawrence Erlbaum and Associates.
- Frazer, N., & Miller, R. (1977). Training practical instructors (programmed patients) to teach basic physical examination. *Journal of Medical Education*, 52(February), 149-151.
- Friedman, M., & Mennin, S. (1991). Rethinking critical issues in performance assessment. *Academic Medicine*, 66(7), 390-395.
- Friedman, R., Korst, D., Schultz, J., Beatty, E., & Entine, S. (1978). Experience with the simulated patient-physician encounter. *Journal of Medical Education*, 53(October), 825-830.
- Furman, G., Colliver, J., & Galofre, A. (1993). Effects of Student Gender and Standardized-patient Gender in a Single Case Using a Male and a Female Standardized Patient. *Academic Medicine*, 68(4), 301-303.
- Furman, G., Colliver, J., Galofre, A., Reaka, M., Robbs, R., & King, A. (1997). The effect of formal feedback sessions on test security for a clinical practice examination using standardized patients. In A. Scherpbier, C. van der Vleuten, J. Rethans, & v. d. S. A (Eds.), *Advances in Medical Education* (pp. 433-436). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Gale, J. (1982). Some cognitive components of the diagnostic thinking process. *British Journal of Educational Psychology*, 52, 64-76.
- Gale, J., & Marsden, P. (1982). Clinical problem solving: the beginning of the process. *Medical Education*, 16(22-26).
- Gandy, J. (1995). Collaboration and Interdependence. *PT Magazine*, February, 40-44.
- Gandy, J. (1999). Preparation for teaching in clinical settings. In K. Shepard & G. Jensen (Eds.), *Handbook of Teaching for Physical Therapists* (pp. 119-167). Boston: Butterworth-Heinemann.
- Gandy, J., & Jensen, G. (1992). Group work and reflective practicums in physical therapy education: models for professional behaviour development. *Journal of Physical Therapy Education*, 6(1), 6-10.
- Garner, R. (1988). Verbal-report data on cognitive and metacognitive strategies. In C. Weinstein, E. Goetz, & P. Alexander (Eds.), *Learning and Study Strategies:*

Issues in Assessment, Instruction and Evaluation (pp. 63-75). New York: Academic Press Inc.

- Garrett, K. (1998). Cooperative learning in social work research courses: helping students to help one another. *Journal of Social Work Education, 34*(2), 237-246.
- Geddes, E., & Crowe, J. (1998). Peer-rated objective structured clinical examination. *Physiotherapy Canada, Fall*, 268-274.
- Gerace, L., & Sibilano, H. (1984). Preparing students for peer collaboration: A clinical teaching model. *Journal of Nursing Education, 23*(5), 206-209.
- Gerber, W., Albanese, M., Brown, D., & Matthes, S. (1985). Teaching with simulated patients: evaluation of the long-term effectiveness of instruction. *Evaluation and the Health Professions, 8*(1), 69-82.
- Gilhooly, K. (1990). Cognitive psychology and medical diagnosis. *Applied Cognitive Psychology, 4*, 261-272.
- Gillette, N., & Mattingly, C. (1987). Clinical reasoning in occupational therapy. *American Journal of Occupational Therapy, 41*, 399-400.
- Gillies, R., & Ashman, A. (1995). Current developments in peer learning. *Unicorn, 21*(1), 77-88.
- Glaser, R., & Chi, M. (1988). Overview. In M. Chi, R. Glaser, & M. Farr (Eds.), *The Nature of Expertise* (pp. xv-xxviii). Hillsdale, New Jersey: Lawrence Erlbaum and Associates.
- Glass, G., McGaw, B., & Smith, M. (1981). *Meta-analysis in Social Research*. London: Sage Publications.
- Gliva, G. (1997). Standardised Patient Protocol, Rotator Cuff Tear : Eastern Virginia Medical School.
- Gold, G., Hadda, C., Taylor, B., Tideiksaar, R., & Mulvihill, M. (1995). A standardised patient program in a mandatory geriatrics clerkship for medical students. *The Gerontologist, 35*(1), 61-66.
- Goldenberg, D., & Iwasiw, C. (1992). Reciprocal learning among students in the clinical area. *Nurse Educator, 17*(5), 27-29.
- Goldschmid, B., & Goldschmid, M. (1976). Peer teaching in higher education: a review. *Higher Education, 5*, 9-33.
- Gomez, J., Prieto, L., Pujol, R., Arbizu, T., Vilar, L., Pi, F., Borrell, F., Roma, J., & Martinez-Carretero, J. (1997). Clinical skills assessment with standardised patients. *Medical Education, 31*, 94-98.
- Gordon, J., Sanson-Fisher, R., & Saunders, N. (1988). Identification of simulated patients by interns in a casualty setting. *Medical Education, 22*, 533-538.
- Graham, C. (1996). Conceptual learning processes in physical therapy students. *Physical therapy, 76*(8), 856-865.

- Grant, J., & Marsden, P. (1987). The Structure of Memorized Knowledge in Students and Clinicians: an Explanation for Diagnostic Expertise. *Medical Education*, 21, 92-98.
- Grant, R., Jones, M., & Maitland, G. (1988). Clinical decision making in upper quadrant dysfunction. In R. Grant (Ed.), *Physical Therapy of the Cervical and Thoracic Spine* (pp. 51-80). New York: Churchill Livingstone.
- Greenwood, C., Carta, J., & Kamps, D. (1990). Teacher-mediated versus peer-mediated instruction: a review of educational advantages and disadvantages. In H. Foot, M. Morgan, & R. Shute (Eds.), *Children Helping Children* (pp. 177-205): John Wiley and Sons Ltd.
- Greenwood, J. (1997). Theoretical approaches to the study of nurses' clinical reasoning: getting things clear. *Contemporary Nurse*, 7, 110-116.
- Greenwood, J., & King, M. (1995). Some surprising similarities in the clinical reasoning of expert and novice orthopaedic nurses: report of a study using verbal protocols and protocol analysis. *Journal of Advanced Nursing*, 22, 907-913.
- Griffin, B., & Griffin, M. (1997). The effects of reciprocal peer tutoring on graduate students' achievement, test anxiety, and academic self-efficacy. *The Journal of Experimental Education*, 65(3), 197-209.
- Griffin, M., & Griffin, B. (1998). An investigation of the effects of reciprocal peer tutoring on achievement, self-efficacy, and test anxiety. *Contemporary Educational Psychology*, 23(3), 298-311.
- Groen, G., & Patel, V. (1988). The relationship between comprehension and reasoning in medical expertise. In M. Chi, R. Glaser, & M. Farr (Eds.), *The Nature of Expertise* (pp. 287-310). Hillsdale, New Jersey: Lawrence Erlbaum and Associates.
- Gronlund, N. (1981). *Measurement and Evaluation in Teaching*. (4 ed.). New York: Macmillan Publishing Company Inc.
- Guba, E., & Lincoln, Y. (1981). *Effective Evaluation: Improving the Effectiveness of Evaluation Results Through Responsive and Naturalistic Approaches*. San Francisco: Jossey-Bass.
- Guralnik, D. (Ed.). (1977). *Webster's New World Dictionary* (2nd Concise Edition ed.). New York: William Collins and World Publishing Co. Inc.
- Haffner-Zavadak, K., Konecky-Dolnack, C., Polich, S., & Van Volkenburg, M. (1995). Collaborative Models. *PT Magazine*, February, 46-54.
- Hagedorn, R. (1996). Clinical decision making in familiar cases: a model of the process and implications for practice. *British Journal of Occupational Therapy*, 59(5), 217-222.
- Hall, R., Rocklin, T., Dansereau, D., Skaggs, L., O'Donnell, A., Lambiotte, J., & Young, M. (1988). The role of individual differences in the cooperative learning of technical material. *Journal of Educational Psychology*, 80(2), 172-178.

- Hampton, D., & Grudnitski, G. (1996). Does cooperative learning mean equal learning? *Journal of Education for Business*, 72(1), 5-7.
- Hanna, D. (1991). Using Simulations to Teach Clinical Nursing. *Nurse Educator*, 16(2), 28-31.
- Hannay, D. (1980). Teaching interviewing with simulated patients. *Medical Education*, 14, 246-248.
- Harden, R., & Gleeson, F. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education*, 13, 41-54.
- Harden, R., Stevenson, M., Downie, W., & Wilson, G. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1, 447-451.
- Harper, A., Roy, W., Norman, G., Rand, C., & Feightner, J. (1983). Difficulties in clinical skills evaluation. *Medical Education*, 17, 24-27.
- Hartley, A. (1990). *Practical Joint Assessment: A sports medicine manual*. St. Louis: Mosby Year Book.
- Hasle, J., Anderson, D., & Szerlip, H. (1994). Analysis of the costs and benefits of using standardised patients to help teach physical diagnosis. *Academic Medicine*, 69(7), 567-570.
- Haydon, R., Donnelly, M., Schwartz, R., Strodel, W., & Jones, R. (1994). Use of Standardized Patients to Identify Deficits in Student Performance and Curriculum Effectiveness. *The American Journal of Surgery*, 168(July), 57-65.
- Hayes, B., & Adams, R. (1995). Parallels between the process of clinical reasoning and categorisation. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 147-156). Oxford, UK: Butterworth Heinemann Ltd.
- Heaton, C., Watson, S., & Alger, E. (1994). Using a standardised patient to teach health appraisal in a problem-based format. *Academic Medicine*, 69(5), 415-416.
- Hekelman, F., Flynn, S., Glover, P., Galazka, S., & Phillips, J. (1994). Peer Coaching in Clinical Teaching. *Evaluation and the Health Professions*, 17(3), 366-381.
- Helfer, R., Black, M., & Teitelbaum, H. (1975). A comparison of pediatric interviewing skills using real and simulated mothers. *Pediatrics*, 55(3), 397-400.
- Henry, J. (1989). Meaning and practice in experiential learning. In S. Warner-Weil & I. McGill (Eds.), *Making Sense of Experiential Learning* (pp. 25-37). Philadelphia: Society for Research into Higher Education and the Open University Press.
- Henry, S., Le Breck, D., & Holzemer, W. (1989). The effect of verbalisation of cognitive processes on clinical decision making. *Research in Nursing and Health*, 12, 187-193.

- Hertzog, C., & Lieble, C. (1996). A study of two techniques for teaching introductory geography: traditional approach versus cooperative learning in the university classroom. *Journal of Geography*, 95(6), 274-280.
- Higgs, J. (1990). Fostering the acquisition of clinical reasoning skills. *New Zealand Journal of Physiotherapy*, December, 13-17.
- Higgs, J. (1992). Developing Clinical Reasoning Competencies. *Physiotherapy*, 78(8), 575-581.
- Higgs, J. (1997). Valuing non-propositional knowledge in clinical reasoning: an educational perspective. *Australian New Zealand Association of Medical Education Bulletin*, 24(1), 7-15.
- Higgs, J., & Edwards, H. (1999). Educating beginning practitioners in the health professions. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners: Challenges for Health Professional Education* (pp. 3-9). Oxford: Butterworth-Heinemann.
- Higgs, J., & Hunt, A. (1999). Rethinking the beginning practitioner: introducing the 'interactional professional'. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners: Challenges for Health Professional Education* (pp. 10-18). Oxford: Butterworth-Heinemann.
- Higgs, J., & Jones, M. (1995). Clinical Reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 3-23). Oxford, UK: Butterworth Heinemann Ltd.
- Higgs, J., & Titchen, A. (1995a). The nature, generation and verification of knowledge. *Physiotherapy*, 81(9), 521-530.
- Higgs, J., & Titchen, A. (1995b). Propositional, professional and personal knowledge in clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 129-146). Oxford, UK: Butterworth-Heinemann Ltd.
- Hill, S., Gay, B., & Topping, K. (1998). Peer-assisted learning beyond school. In K. Topping & S. Ehly (Eds.), *Peer Assisted Learning* (pp. 291-311). London: Lawrence Erlbaum and Associates.
- Hilliard, R., & Tallett, S. (1998). The use of an objective structured clinical examination with postgraduate residents in pediatrics. *Archives of Pediatric Adolescent Medicine*, 152, 74-78.
- Hobus, P., Schmidt, H., Boshuizen, H., & Patel, V. (1987). Contextual factors in the activation of first diagnostic hypotheses: expert-novice differences. *Medical Education*, 21, 471-476.
- Hodges, B., Regehr, G., Hanson, M., & McNaughton, N. (1998). Validation of an objective structured clinical examination in psychiatry. *Academic Medicine*, 73(8), 910-912.
- Holt, D., Michael, S., & Godfrey, J. (1997). The case against cooperative learning. *Issues in Accounting Education*, 12(1), 191-193.
- Holzemer, W., Resnik, B., & Slichter, M. (1986). Criterion-related validity of a clinical simulation. *Journal of Nursing Education*, 25(7), 286-290.

- Hunt, A., & Higgs, J. (1999). Learning generic skills. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners: Challenges for Health Professional Education* (pp. 166-172). Oxford: Butterworth-Heinemann.
- Isles, R. (1995). *Clinical decision making in physiotherapy students with neurological patients*. Paper presented at the Proceedings of the 12th International Congress of the World Confederation for Physical Therapy, Washington DC.
- Itano, J. (1989). A comparison of the clinical judgment process in experienced registered nurses and student nurses. *Journal of Nursing Education*, 28(3), 120-125.
- Iwasiw, C., & Goldenberg, D. (1993). Peer teaching among nursing students in the clinical area: effects on student learning. *Journal of Advanced Nursing*, 18, 659-668.
- Jain, S., Nadler, S., Eyles, M., Kirshblum, S., DeLisa, J., & Smith, A. (1997). Development of an objective structured clinical examination (OSCE) for physical medicine and rehabilitation residents. *American Journal of Physical Medicine and Rehabilitation*, 72(2), 102-106.
- Jenkins, P. (1991). The use of actors as simulated patients - three years experience on a day release course for GP trainees. *Postgraduate education for general practice*, 2, 16-27.
- Jennett, P., Tambay, J., Atkinson, M., Baumber, J., Crutcher, R., Hogan, D., Elford, R., MacCannell, K., & Swanson, R. (1992). Chart stimulated recall: a method for assessing factors which influence physicians' practice performance. In R. Harden & I. Hart (Eds.), *Approaches to the Assessment of Clinical Competence* (pp. 511-517). Dundee, Scotland: Page Brothers.
- Jensen, G., Shepard, K., & Hack, L. (1990). The novice versus the experienced clinician: insights into the work of the physical therapist. *Physical Therapy*, 70(5), 314-323.
- Jensen, G., Shepard, K., Gwyer, J., & Hack, L. (1992). Attribute Dimensions that Distinguish Master and Novice Physical Therapy Clinicians in Orthopedic Settings. *Physical Therapy*, 72(10), 711-722.
- Jette, A. (1989). Diagnosis and classification by physical therapists: a special communication. *Physical Therapy*, 69(11), 967-969.
- Johnson, D. (1981). Student-student interaction: the neglected variable in education. *Educational Researcher*, 1, 5-10.
- Johnson, D., & Johnson, R. (1978). Cooperative, Competitive, and Individualistic Learning. *Journal of Research and Development in Education*, 12(1), 3-15.
- Johnson, D., & Johnson, R. (1987). Research Shows the Benefits of Adult Cooperation. *Educational Leadership*, 45, 27-30.
- Johnson, D., & Johnson, R. (1991). *Learning Together and Alone: Cooperative, Competitive and Individualistic Learning*. (3rd ed.). Boston: Allyn and Bacon.

- Johnson, D., Johnson, R., & Smith, K. (1986). Academic conflict among students: controversy and learning. In R. Feldman (Ed.), *The Social Psychology of Education* (pp. 199-229). Cambridge: Cambridge University Press.
- Johnson, D., Johnson, R., & Smith, K. (1998). Cooperative learning returns to college: what evidence is there that it works? *Change*, *30*(4), 27-35.
- Johnson, D., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. (1981). Effects of Cooperative, Competitive, and Individualistic Goal Structures on Achievement: A Meta-Analysis. *Psychological Bulletin*, *89*(1), 47-62.
- Jolly, B. (1982). A review of issues in live patient simulation. *Programmed Learning and Educational Technology*, *19*(2), 99-107.
- Jolly, B., Jones, A., Dacre, J., Elzubeir, M., Kopelman, P., & Hitman, G. (1996). Relationships between student's clinical experiences in introductory clinical courses and their performance on an objective structured clinical examination (OSCE). *Academic Medicine*, *71*(8), 909-916.
- Jones, J. (1988). Clinical reasoning in nursing. *Journal of Advanced Nursing*, *13*, 185-192.
- Jones, J. (1989). The verbal protocol: a research technique for nursing. *Journal of Advanced Nursing*, *14*, 1062-1070.
- Jones, M. (1992). Clinical Reasoning in Manual Therapy. *Physical Therapy*, *72*(12), 875-884.
- Jones, M. (1995). Clinical Reasoning and Pain. *Manual Therapy*, *1*, 17-24.
- Jones, M. (1997). Clinical reasoning: the foundation of clinical practice Part 1 and 2. *Australian Physiotherapy Journal*, *43*(3), 167-170, 213-217.
- Jones, M., Jensen, G., & Edwards, I. (In press). Clinical Reasoning in Physiotherapy. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed.,). Oxford, UK: Butterworth-Heinemann.
- Jones, M., Jenson, G., & Rothstein, J. (1995). Clinical reasoning in physiotherapy. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 72-87). Oxford, UK: Butterworth Heinemann Ltd.
- Joseph, G., & Patel, V. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making*, *10*(1), 31-46.
- Joyce, B., & Showers, B. (1982). The coaching of teaching. *Educational Leadership*, *40*(1), 4-8.
- Joyce, B., & Showers, B. (1995). *Student Achievement Through Staff Development*. (2nd ed.). White Plains, N.Y.: Longman Publishers.
- Joyce, B., & Weil, M. (1996). *Models of Teaching*. (5th ed.). Boston: Allyn and Baron.
- Joyce, B., & Weil, M. (1996). The construction of knowledge, metacognition, and conceptions of intelligence, *Models of Teaching* (5th ed.,). Boston: Allan and Bacon.

- Kagan, N., & Krathwohl, D. (1967). *Studies in Human Interaction: Interpersonal Process Recall Stimulated by Videotape*. (5-0800): U.S. Department of Health, Education and Welfare.
- Kaiser, S., & Bauer, J. (1995). Checklist self-evaluation in a standardised patient exercise. *The American Journal of Surgery*, 169(April), 418-420.
- Kaufman, R., Portney, L., & Jette, D. (1997). Clinical performance of physical therapy students in traditional and problem based curricula. *Journal of Physical Therapy Education*, 11(1), 26-31.
- Kazdin, A. (1982). *Single-Case Research Designs*. New York: Oxford University Press.
- Kiess, H. (1996). *Statistical Concepts for the Behavioural Sciences*. (2nd ed.). Boston: Allyn and Bacon.
- King, A. (1997). ASK to THINK-TEL WHY: A model of transactive peer tutoring for scaffolding higher level complex learning. *Educational Psychologist*, 32(4), 221-235.
- Kinnersley, P., & Pill, R. (1993). Potential of using simulated patients to study the performance of general practitioners. *British Journal of General Practice*, 43(372), 297-300.
- Klass, D., Campbell, C., Hassard, T., Kopelow, M., & Schnabl, G. (1990). *Influence of level of training on performance in a standardised test of clinical abilities*. Paper presented at the Teaching and Assessing Clinical Competence, Groningen, The Netherlands.
- Kleffner, J., & Dadian, T. (1997). Using collaborative learning in dental education. *Journal of Dental Education*, 61(1), 66-72.
- Kohler, F., McCullough Crilley, K., & Shearer, D. (1997). Effects of peer coaching on teacher and student outcomes. *The Journal of Educational Research*, 90(4), 240-250.
- Kopelow, M., Schnabl, G., Hassard, T., Tamblyn, R., Klass, D., Beazley, G., Hechter, F., & Grott, M. (1992). Assessment of performance in the office setting with standardised patients. *Academic Medicine*, 67(10), S19-S21.
- Kuketich, G., Colliver, J., Galofre, A. (1992). Two studies of the effects of gender within a single case simulated by a male and female standardised patient in a multiple-station examination. In R. Harden, I. Hart & H. Mulholland (Eds.), *Approaches to the Assessment of Clinical Competence* (pp. 199-203). Dundee, Scotland: Page Brothers.
- Ladyshevsky, R. (1993). Clinical teaching and the 2:1 student-to-clinical instructor ratio. *Journal of Physical Therapy Education*, 7(1), 31-35.
- Ladyshevsky, R. (1996). East meets West: the influence of language and culture in clinical education. *Australian Physiotherapy Journal*, 42(4), 287-294.
- Ladyshevsky, R., & Gotjamanos, E. (1997). Communication skill development in health professional education: The use of standardised patients in combination with a peer assessment strategy. *Journal of Allied Health*, 26(4), 177-186.

- Ladyshefsky, R., & Healey, E. (1990). *The 2:1 Teaching Model in Clinical Education. A Manual for Clinical Instructors*. Toronto: University of Toronto, Division of Physical Therapy.
- Ladyshefsky, R., Barrie, S., & Drake, V. (1998). A comparison of productivity and learning outcome in individual and cooperative physical therapy clinical education models. *Physical Therapy, 78*(12), 1288-1301.
- Lake, D. (1999). Enhancement of student performance in a gross anatomy course with the use of peer tutoring. *Journal of Physical Therapy Education, 13*(1), 34-38.
- Larkin, J., McDermott, J., Simon, D., & Simon, H. (1980). Expert and novice performance in solving physics problems. *Science, 208*(June), 1335-1342.
- Larson, C., Dansereau, D., O'Donnell, A., Hythecker, V., Lambiotte, J., & Rocklin, T. (1985). Effects of Metacognitive and Elaborative Activity on Cooperative Learning and Transfer. *Contemporary Educational Psychology, 10*, 342-348.
- Leighton, R., & Sheldon, M. (1997). Model for Teaching Clinical Decision Making in a Physical Therapy Professional Curriculum. *Journal of Physical Therapy Education, 11*(2), 23-30.
- Lemke, T., & Basile, C. (1997). An odyssey into cooperative learning. *American Journal of Pharmaceutical Education, 61*, 351-358.
- Lincoln, M., & McAllister, L. (1993). Peer learning in clinical education. *Medical Teacher, 15*(1), 17-25.
- Lincoln, Y., & Guba, E. (1985). *Naturalistic Inquiry*. Beverly Hills: Sage.
- Lindquist, T. (1997). An experimental test of cooperative learning with faculty members as subjects. *Journal of Education for Business, 3*, 157-163.
- Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-24.
- Liu, L., Schneider, P., & Miyazaki, M. (1997). The effectiveness of using simulated patients versus videotapes of simulated patients to teach clinical skills to occupational and physical therapy students. *The Occupational Therapy Journal of Research, 17*(3), 159-172.
- Liu, P., Miller, E., Herr, G., Hardy, C., Sivarajan, M., & Willenkin, R. (1980). Videotape reliability: A method of evaluation of a clinical performance examination. *Journal of Medical Education, 55*(August), 713-715.
- Lloyd, J., Williams, R., Simonton, D., & Sherman, D. (1990). Order effects in standardized patient examinations. *Academic Medicine, 65*(9), S51-S52.
- Lynch, B. (1984). Cooperative Learning in Interdisciplinary Education for the Allied Health Professions. *Journal of Allied Health, 13*(2), 83-93.
- MacRae, H., Vu, N., Graham, B., Word-Sims, M., Colliver, J., & Robbs, R. (1995). Comparing checklists and databases with physicians' ratings as measures of students' history and physical-examination skills. *Academic Medicine, 70*(4), 313-317.

- Maheady, L. (1998). Advantages and disadvantages of peer assisted learning strategies. In K. Topping & S. Ehly (Eds.), *Peer Assisted Learning* (pp. 45-65). London: Lawrence Erlbaum and Associates.
- Maitland, G. (1977). *Peripheral Manipulation*. (Second ed.). London: Butterworth and Co (Publishers) Ltd.
- Martin, G., & Double, J. (1998). Developing higher education teaching skills through peer observation and collaborative reflection. *Innovations in Education and Training International*, 35(2), 161-169.
- Martin, M., & Edwards, L. (1998). Peer learning on fieldwork placements. *British Journal of Occupational Therapy*, 61(6), 249-252.
- Mast, T., Feltovitch, P., Soler, N., & Myers, A. (1985). *The assessment of medical clerk's understanding using structured simulated recall*. Paper presented at the New Developments in Clinical Competence, Ottawa, Canada.
- Mattana, D., Shepherd, K., & Knight, G. (1997). The effect of cooperative learning strategy on preclinical performance. *Journal of Dental Education*, 61(6), 480-483.
- Mattingly, C. (1991a). The narrative nature of clinical reasoning. *American Journal of Occupational Therapy*, 45, 998-1005.
- Mattingly, C. (1991b). What is clinical reasoning? *American Journal of Occupational Therapy*, 45, 979-986.
- Maudsley, G. (1999). Do we all mean the same thing by "problem-based learning"? A review of the concepts and a formulation of the ground rules. *Academic Medicine*, 74(2), 178-184.
- May, B., & Newman, J. (1980). Developing competence in problem solving. *Physical Therapy*, 60(9), 1140-1145.
- McAvoy, B. (1988). Teaching clinical skills to medical students: the use of simulated patients and videotaping in general practice. *Medical Education*, 22, 193-199.
- McClure, C., Gall, E., Meredith, K., Gooden, M., & Boyer, J. (1985). Assessing clinical judgement with standardised patients. *The Journal of Family Practice*, 20(5), 457-464.
- McDowell, J., Nardini, D., Negley, S., & White, J. (1984). Evaluating clinical performance using simulated patients. *Journal of Nursing Education*, 23(1), 37-39.
- McKay, E., & Ryan, S. (1995). Clinical reasoning through story telling: examining a student's case story on a fieldwork placement. *British Journal of Occupational Therapy*, 58(6), 234-238.
- McKnight, J., Rideout, E., Brown, B., Ciliska, D., Patton, D., Rankin, J., & Woodward, C. (1987). The objective structured clinical examination: An alternative approach to assessing student clinical performance. *Journal of Nursing Education*, 26(1), 39-41.

- McLaughlin, T., Mecham, M., & Montague, P. (1995). Peer tutoring: history, effectiveness, and lack of wide spread implementation. *Corrective and Social Psychiatry Journal*, 41(4), 71-78.
- Mezirow, J. (1981). A critical theory of adult learning and education. *Adult Education*, 31(1), 3-24.
- Miles, M., & Huberman, A. (1994). *Qualitative Data Analysis*. (2nd ed.). London: Sage Publications.
- Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Miller, G. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, 65(9 September Supplement), S63-S67.
- Milson, L., & Laatsch, L. (1996). Does cooperative learning enhance student achievement? *Laboratory Medicine*, 27(9), 618-621.
- Monaghan, M., Vanderbush, R., Gardner, S., Schneider, E., Grady, A., & McKay, A. (1997). Standardized patients: an ability-based outcomes assessment for the evaluation of clinical skills in traditional and nontraditional education. *American Journal of Pharmaceutical Education*, 61, 337-344.
- Morrison, L. (1997). Director, Standardised Patient Program : Southern Illinois University of Medicine.
- Moss, P. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research*, 62(3), 229-258.
- Mumford, E., Anderson, D., Cuerdon, T., & Scully, J. (1984). Performance-based evaluation of medical students' interviewing skills. *Journal of Medical Education*, 59(February), 133.
- Muzzin, L., Norman, G., Feightner, J., Tugwell, P., & Guyatt, G. (1983). *Expertise in recall of clinical protocols in two specialty areas*. Paper presented at the 22nd Annual Conference on Research in Medical Education.
- Muzzin, L., Norman, G., Jacoby, L., Feightner, J., Tugwell, P., & Guyatt, G. (1982). *Manifestations of expertise in recall of clinical protocols*. Paper presented at the 21st Annual Conference on Research in Medical Education, Washington, D.C.
- Nayer, M. (1993). An overview of the objective structured clinical examination. *Physiotherapy Canada*, 45(3), 171-178.
- Nayer, M. (1995). The assessment of clinical competency: An overview and preliminary report of Canadian physiotherapy programs. *Physiotherapy Canada*, 47(3), 190-199.
- Neistadt, M. (1996). Teaching strategies for the development of clinical reasoning. *American Journal of Occupational Therapy*, 50(8), 676-684.
- Neistadt, M., & Smith, R. (1997). Teaching diagnostic reasoning: using a classroom as clinic methodology with videotapes. *American Journal of Occupational Therapy*, 51(5), 360-368.

- Nelson, L. (1981). *Platista: An Introduction to Applied Social Science Statistical Methods*. Dunedin, New Zealand: Department of Education, University of Otago.
- Neufeld, V., Norman, G., Feightner, J., & Barrows, H. (1981). Clinical problem solving by medical students: a cross-sectional and longitudinal analysis. *Medical Education, 15*, 315-322.
- Newble, D., Baxter, A., & Elmslie, R. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education, 13*, 263-268.
- Newble, D., Hoare, J., & Elmslie, R. (1981). The validity and reliability of a new examination of the clinical competence of medical students. *Medical Education, 15*, 46-52.
- Newble, D., van der Vleuten, C., & Norman, G. (1995). Assessing clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 168-178). Oxford, UK: Butterworth Heinemann Ltd.
- Nickerson, R., Perkins, D., & Smith, E. (1985). *The Teaching of Thinking*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Nisbett, R., & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, New Jersey: Prentice-Hall.
- Nisbett, R., & Wilson, T. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84*, 231-259.
- Norcini, J. (1992). *Approaches to standard setting for performance-based examinations*. Paper presented at the Approaches to the Assessment of Clinical Competence, Dundee, Scotland.
- Norman, G. (1985). *Evaluation of problem-solving ability*. Paper presented at the New Developments in Clinical Competence, Ottawa, Canada.
- Norman, G., & Schmidt, H. (1992). The psychological basis of problem-based learning: a review of the evidence. *Academic Medicine, 67*(9), 557-565.
- Norman, G., Barrows, H., Gliva, G., & Woodward, C. (1985a). Simulated Patients. In V. Neufeld & G. Norman (Eds.), *Assessing Clinical Competence* (pp. 219-228). New York, New York: Springer Publishing Company.
- Norman, G., Coblenz, C., Brooks, L., & Babcock, C. (1992). Expertise in visual diagnosis: a review of the literature. *Academic Medicine, 67*, S78-S83.
- Norman, G., Davis, D., Painvin, A., Lindsay, E., Rath, D., & Ragbeer, M. (1989). *Comprehensive assessment of clinical competence of family/general physicians using multiple measures*. Paper presented at the Twenty-eighth Annual Conference of Research in Medical Education, Washington, D.C.
- Norman, G., Feightner, J., & Tugwell, P. (1983). *The generalisability of measures of clinical problem solving*. Paper presented at the Twenty Second Conference on Research in Medical Education, Washington, D.C.

- Norman, G., Jacoby, L., Feightner, J., & Campbell, E. (1979). *Clinical experience and the structure of memory*. Paper presented at the 18th Annual Conference on Research in Medical Education, Washington, D.C.
- Norman, G., Muzzin, L., Williams, R., & Swanson, D. (1985). Simulation in Health Sciences Education. *Journal of Instructional Development*, 8(1), 11-17.
- Norman, G., Neufeld, V., Walsh, A., Woodward, C., & McConvey, G. (1985c). Measuring Physicians' Performances By Using Simulated Patients. *Journal of Medical Education*, 60, 925-934.
- Norman, G., Tugwell, P., & Feightner, J. (1982). A Comparison of Resident Performance On Real and Simulated Patients. *Journal of Medical Education*, 57(September), 708-715.
- Norman, G., Tugwell, P., Feightner, J., Muzzin, L., & Jacoby, L. (1985). Knowledge and clinical problem solving. *Medical Education*, 19, 344-356.
- Norman, G., van der Vleuten, C., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of validity, efficiency and acceptability. *Medical Education*, 25, 119-126.
- Norton, B., & Strube, M. (1998). The influence of experience with a set of simulated patients on diagnosis of simulated patients not previously diagnosed. *Physical Therapy*, 78(4), 375-385.
- Nowotny, R., & Grove, D. (1982). Description of an examination for the objective assessment of history-taking ability. *Medical Education*, 16, 259-263.
- O'Donnell, A., & Topping, K. (1998). Peers assessing peers: possibilities and problems. In K. Topping & S. Ehly (Eds.), *Peer Assisted Learning* (pp. 255-278). London: Lawrence Erlbaum and Associates.
- O'Donnell, A., Dansereau, D., Rocklin, T., Hythecker, V., Lambiotte, J., Larson, C., & Young, M. (1985). Effects of elaboration frequency on cooperative learning. *Journal of Educational Psychology*, 77(5), 572-580.
- O'Hagan, J., Davies, L., & Pears, R. (1986). The use of simulated patients in the assessment of actual clinical performance in general practice. *New Zealand Medical Journal*, 99, 948-951.
- Ogden, A., Barnhart, A., & Davis, P. (1987). *The extended standardised patient as an assessment and teaching tool*. Paper presented at the Further Developments in Assessing Clinical Competence, Ottawa, Canada.
- Oldmeadow, L. (1996). Developing clinical competence: a mastery pathway. *Australian Physiotherapy Journal*, 42(1), 37-44.
- Oppenheim, A. (1992). *Questionnaire Design, Interviewing and Attitude Measurement*. (New ed.). London: Pinter Publishers.
- Ostrow, D. (1980). Surrogate patients in medical education. *Programmed Learning Educational Technology*, 17(2), 82-89.
- Owen, A., & Underwood, P. (1980). Videotape and simulated patients. *Medical Journal of Australia*, 1(May), 437-440.

- Owen, A., & Winkler, R. (1974). General practitioners and psychosocial problems: an evaluation using pseudopatients. *Medical Journal of Australia*, 2(September), 393-398.
- Page, G., & Bordage, G. (1995). The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Academic Medicine*, 70(2), 104-110.
- Page, G., Bordage, G., & Allen, T. (1995). Developing key-feature problems and examinations to assess clinical decision-making skills. *Academic Medicine*, 70(3), 194-201.
- Palmer, M., & Epler, M. (1990). *Clinical Assessment Procedures in Physical Therapy*. Philadelphia: J.B. Lippincott Company.
- Papa, F., Shores, J., Meyers, S., O'Reilly, R., & Bourdage, R. (1990a). The role of pattern matching and pattern differentiation in clinical problem solving. In W. Bender, R. Hiemstra, A. Scherpbier, & R. Zwierstra (Eds.), *Teaching and Assessing Clinical Competence* (pp. 317-322). Groningen, Netherlands: Boerkerk.
- Papa, F., Shores, T., & Mayer, S. (1990b). Effects of pattern matching, pattern discrimination and experience in the development of diagnostic expertise. *Academic Medicine*, 65(9), S21-S22.
- Paris, S. (1985). Clinical decision making: orthopedic physical therapy. In S. Wolf (Ed.), *Clinical Decision Making in Physical Therapy* (pp. 215-253). Philadelphia: F.A. Davis Company.
- Parker-Taillon, D., Cornwall, J., Cohen, R., & Rothman, A. (Eds.). (1992). *The development of a physiotherapy national examination OSCE*. (Vol. 1). Dundee, Scotland: Page Brothers.
- Paschal, K., & Jensen, G. (1995). *Developing expertise: A model for clinical education*. Paper presented at the Proceedings of the 12th International Congress of the World Confederation for Physical Therapy, Washington DC.
- Patel, V., & Arocha, J. (1995). Methods in the study of clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 35-48). Oxford: Butterworth-Heinemann Ltd.
- Patel, V., & Coughlin, L. (1985). *Cognitive science and the assessment of clinical competence*. Paper presented at the New Developments in Clinical Competence.
- Patel, V., & Groen, G. (1986a). Differences between medical students and doctors in memory for clinical cases. *Medical Education*, 20, 3-9.
- Patel, V., & Groen, G. (1986b). Knowledge-based solution strategies in medical reasoning. *Cognitive Science*, 10, 91-116.
- Patel, V., & Kaufman, D. (1995). Clinical reasoning and biomedical knowledge: implications for teaching. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning and the Health Professions* (pp. 117-128). Oxford, UK: Butterworth Heinemann Ltd.

- Patel, V., Evans, D., & Groen, G. (1989). On reconciling basic science and clinical reasoning. *Teaching and Learning in Medicine: An International Journal*, 1, 116-121.
- Patel, V., Groen, G., & Arocha, J. (1990). Medical expertise as a function of task difficulty. *Memory and Cognition*, 18, 394-406.
- Patel, V., Groen, G., & Scott, H. (1988). Biomedical knowledge in explanations of clinical problems by medical students. *Medical Education*, 22, 398-406.
- Payton, O. (1985). Clinical Reasoning Process in Physical Therapy. *Physical Therapy*, 65(6), 924-928.
- Perkowski-Rogers, L., Solomon, D., Speer, A., & Ainsworth, M. (1992). *Clinical relevance of station length and checklist scores on the assessment of clinical skills*. Paper presented at the Approaches to the Assessment of Clinical Competence, Dundee, Scotland.
- Perry, M. (1988). Preceptorship in clinical nursing education: a social learning theory approach. *The Australian Journal of Advanced Nursing*, 5(3), 19-25.
- Pesut, D., & Herman, J. (1992). Metacognitive skills in diagnostic reasoning: making the implicit explicit. *Nursing Diagnosis*, 3, 148-154.
- Peterson, P., & Clark, C. (1978). Teachers' reports of their cognitive processes during teaching. *American Educational Research Journal*, 15, 555-565.
- Petrusa, E., Blackwell, T., Rogers, L., Saydjari, C., Parcel, S., & Guckian, J. (1987). An objective measure of clinical performance. *American Journal of Medicine*, 83, 34-42.
- Piaget, J. (1977). *The Moral Judgment of the Child*. London: Penguin.
- Pickard, C., Potter, H., Boyle, J., & Owens, J. (1996). *Orthopaedic Science 252 Manual*. Perth, Western Australia: School of Physiotherapy, Curtin University of Technology.
- Potter, H. (1995). *Manipulative Therapy: A Guide to Assessment and Treatment of Peripheral Joints*. Perth, Western Australia: School of Physiotherapy, Curtin University of Technology.
- Potter, H., Pickard, C., Williams, L., & Liston, C. (1996). *Orthopaedic Science 351 Manual*. Perth, Western Australia: School of Physiotherapy, Curtin University of Technology.
- Qin, Z., Johnson, D., & Johnson, R. (1995). Cooperative versus competitive efforts and problem solving. *Review of Educational Research*, 65(2), 129-143.
- Quellmalz, E. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education*, 4(4), 319-331.
- Ramsden, P. (1985). Student learning research: retrospect and prospect. *Higher Education Research and Development*, 4(1), 51-69.
- Ramsden, P. (1988). Studying Learning: Improving Teaching. In P. Ramsden (Ed.), *Improving Learning: New Perspectives* (pp. 13-31). London: Kogan Page.

- Ravenscroft, S. (1997). In support of cooperative learning. *Issues in Accounting Education, 12*(1), 187-190.
- Refshauge, K., & Higgs, J. (1995). Teaching clinical reasoning in health science curricula. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 105-116). Oxford, UK: Butterworth Heinemann Ltd.
- Regehr, G., & Norman, G. (1996). Issues in cognitive psychology: implications for professional education. *Academic Medicine, 71*(9), 988-1000.
- Regehr, G., MacRae, H., Reznick, R., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine, 73*(9), 993-997.
- Resnick, L. (1988). Learning in school and out. *Educational Researcher, 16*(9), 13-20.
- Retegui, J., & Cornel-Avendano, B. (1999). *Mastering the OSCE/CSA*. New York: McGraw Hill.
- Rethans, J., Drop, R., Sturmans, F., & van der Vleuten, C. (1991). A method for introducing standardized (simulated) patients into general practice consultations. *British Journal of General Practice, 41*(March), 94-96.
- Rethans, J., Sturmans, F., Drop, R., & van der Vleuten, C. (1991a). Assessment of the performance of general practitioners by the use of standardized (simulated) patients. *British Journal of General Practice, 41*(March), 97-99.
- Rethans, J., Sturmans, F., Drop, R., van der Vleuten, C., & Hobus, P. (1991b). Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *British Medical Journal, 303*(November), 1377-1380.
- Rethans, J., van Leeuwen, Y., Drop, R., van der Vleuten, C., & Sturmans, F. (1990). Competence and performance: Two different concepts in the assessment of quality of medical care. *Family Practice, 7*(3), 168-174.
- Riggio, R., Fantuzzo, J., Connelly, S., & Dimeff, L. (1991). Reciprocal Peer Tutoring: A classroom strategy for promoting academic and social integration in undergraduate students. *Journal of Social Behaviour and Personality, 6*(2), 387-396.
- Riggio, R., Whatley, M., & Neale, P. (1994). Effects of Student Academic Ability on Cognitive Gains Using Reciprocal Peer Tutoring. *Journal of Social Behaviour and Psychology, 9*(3), 529-542.
- Ripkey, D., Case, S., & Swanson, D. (1996). A "new" item format for assessing aspects of clinical competence. *Academic Medicine, 71*(10), S34-S36.
- Rivett, D., & Higgs, J. (1997). Hypothesis generation in the clinical reasoning behaviour of manual therapists. *Journal of Physical Therapy Education, 11*(1), 40-45.

- Robb, K., & Rothman, A. (1985). The assessment of clinical skills in general medical residents - comparison of the objective structured clinical examination to a conventional oral examination. *Annals RCPSC, 18*(3), 235-238.
- Roberts, A. (1996). Approaches to reasoning in occupational therapy: a critical exploration. *British Journal of Occupational Therapy, 59*(5), 233-236.
- Rogers, J., & Holm, M. (1991). Occupational therapy diagnostic reasoning: a component of clinical reasoning. *American Journal of Occupational Therapy, 45*, 1045-1053.
- Rosebraugh, C., Speer, A., Ainsworth, M., Solomon, D., Callaway, M., & Holden, M. (1996). Developing a presentation and problem-solving station in a multistation standardized-patient examination. *Academic Medicine, 71*(1), s102-s104.
- Rothman, A., & Cohen, R. (1995). Understanding the Objective Structured Clinical Examination: Issues and Options. *Annals RCPSC, 28*(5), 283-287.
- Rothstein, J., & Echternach, J. (1986). Hypothesis-oriented algorithm for clinicians: a method for evaluation and treatment planning. *Physical Therapy, 66*(9), 1388-1394.
- Rutala, P., Witzke, D., Leko, E., & Fulginiti, J. (1991a). The Influences of student and standardised patient genders on scoring in an objective structured clinical examination. *Academic Medicine, 66*(9), S28-S30.
- Rutala, P., Witzke, D., Leko, E., Fulginiti, J., & Taylor, P. (1991b). Sharing of information by students in an objective structured clinical examination. *Arch Intern Med, 151*(March), 541-544.
- Saarinen-Rahiiika, H., & Binkley, J. (1998). Problem-based learning in physical therapy: a review of the literature and overview of the McMaster University experience. *Physical Therapy, 78*(2), 195-207.
- Sahrman, S. (1988). Diagnosis by the physical therapist-a prerequisite for treatment. *Physical Therapy, 68*(11), 1703-1706.
- Sanson-Fisher, R., & Poole, A. (1980). Simulated patients and the assessment of medical students' interpersonal skills. *Medical Education, 14*, 249-253.
- Schell, B., & Cervero, R. (1993). Clinical reasoning in occupational therapy: An integrative review. *American Journal of Occupational Therapy, 47*(7), 605-610.
- Schmidt, H. (1983). Problem based learning: rationale and description. *Medical Education, 17*, 11-16.
- Schmidt, H., & Boshuizen, H. (1993). On acquiring expertise in medicine. *Educational Psychology Review, 5*(3), 205-221.
- Schmidt, H., Norman, G., & Boshuizen, H. (1990). A cognitive perspective on medical expertise: theory and implications. *Academic Medicine, 65*(10), 611-621.
- Schon, D. (1983). *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books Inc.

- Schon, D. (1991). *The Reflective Practitioner: How Professionals Think in Action*. London: Ashgate Publishing Ltd.
- School of Physiotherapy, Curtin University of Technology. (1996). Orthopaedic Science 351 Unit Outline . Perth, Western Australia.
- School of Physiotherapy, Curtin University of Technology. (1997). Orthopaedic Science 252 Unit Outline . Perth, Western Australia.
- Schunk, D. (1998). Peer modeling. In K. Topping & S. Ehly (Eds.), *Peer Assisted Learning* (pp. 185-202). London: Lawrence Erlbaum and Associates.
- Scott, I. (1996). Understanding and developing clinical reasoning skills. *Australasian and New Zealand Association for Medical Education: Occasional Paper, 1*, 1-40.
- Scott, T. (1994). *Collaborative Learning*. Paper presented at the 6th Ottawa Conference on Medical Education, Toronto.
- Sharpley, A., & Sharpley, C. (1981). Peer tutoring: A review of the literature. *Collected Original Resources in Education, 5*(3), 1-63.
- Shatzer, J., Wardrop, J., Bible, K., Williams, R., Rhee, K., Carlson, M., & Witkovsky, M. (1990). *Station length requirements for SP exams: their influence on generalisability estimates*. Paper presented at the Current Developments in Assessing Clinical Competence, Ottawa, Canada.
- Shepard, K., & Jensen, G. (1999). Preparation for teaching in academic settings. In K. Shepard & G. Jensen (Eds.), *Handbook of Teaching for Physical Therapists* (pp. 37-72). Boston: Butterworth-Heinemann.
- Showers, B. (1984). *Peer Coaching: A Strategy for Facilitating Transfer of Training* : Oregon University Center for Educational Policy and Management.
- Showers, B. (1985). Teachers coaching teachers. *Educational Leadership, 42*(7), 43-48.
- Siegel, L., Siegel, L., Capretta, P., Jones, R., & Berkowitz, H. (1963). Students' thoughts during class: a criterion for educational research. *Journal of Educational Psychology, 54*, 45-51.
- Skinner, B., Newton, W., & Curtis, P. (1997). The educational value of an OSCE in a family practice residency. *Academic Medicine, 72*(8), 722-724.
- Skinner, M., & Welch, T. (1996). Peer coaching for better teaching. *College Teaching, 44*(4), 153-156.
- Slavin, R. (1977). Classroom reward structure: analytical and practical review. *Review of Educational Research, 47*, 633-650.
- Slavin, R. (1983a). *Cooperative Learning*. New York: Longman Inc.
- Slavin, R. (1983b). When Does Cooperative Learning Increase Student Achievement? *Psychological Bulletin, 94*(3), 429-445.
- Slavin, R. (1987). Developmental and motivational perspectives on cooperative learning: A reconciliation. *Child Development, 58*, 1161-1167.

- Slavin, R. (1990). Research on cooperative learning: consensus and controversy. *Educational Leadership*, 47(4), 52-54.
- Slavin, R. (1991). Group rewards make groupwork work. *Educational Leadership*, 48(5), 89-91.
- Slavin, R. (1995). *Cooperative Learning: Theory, Research and Practice*. (2nd ed.). Boston: Allyn and Bacon.
- Sloan, D., Bonnelly, M., & Schwartz, M. (1993). Student Performance on Examination of Two Simulated Patients with Abdominal Pain. *Annals RCPSC*, 26(3), 171-174.
- Sloan, D., Donnelly, M., Schwartz, R., Munch, L., & Wells, M. (1994). Assessing medical students' and surgery residents' clinical competence in problem solving in surgical oncology. *Annals of Surgical Oncology*, 1(3), 204-212.
- Smit, G., & Van Der Molden, H. (1996). Simulations for the assessment of counselling skills. *Assessment and Evaluation in Higher Education*, 21(4), 335-345.
- Smith, E., & Miller, F. (1978). Limits on perception of cognitive processes: a reply to Nisbett and Wilson. *Psychological Review*, 85, 355-362.
- Smith, M., Hinckley, C., & Volk, G. (1991). Cooperative learning in the undergraduate laboratory. *Journal of Chemical Education*, 68(5), 413-415.
- Soeters, D., Scherpbier, A., & van Lunsen, H. (1987). Assessing students' performances or... examination peculiarities. In I. Hart & R. Harden (Eds.), *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal Publications.
- Sogunro, O. (1998). Impact of evaluation anxiety on adult learning. *Journal of Research and Development in Education*, 31(2), 109-121.
- Solomon, D., Speer, A., Perkowski, L., & DiPette, D. (1994). Evaluating Problem Solving Based on the Use of History Findings in a Standardized Patient Examination. *Academic Medicine*, 69(9), 754-757.
- Stillman, P. (1980). Arizona Clinical Interview Medical Rating Scale. *Medical Teacher*, 2(5), 248-251.
- Stillman, P. (1990). *Standardising standardised patients can it be done?* Paper presented at the Current Developments in Assessing Clinical Competence, Ottawa, Canada.
- Stillman, P. (1993). Technical Issues: Logistics. *Academic Medicine*, 68(6), 464-468.
- Stillman, P., & Swanson, D. (1987). Ensuring the clinical competence of medical school graduates through standardised patients. *Archives of Internal Medicine*, 147(June), 1049-1052.
- Stillman, P., Brown, D., Redfield, D., & Sabers, D. (1977). Construct validation of the Arizona clinical interview rating scale. *Educational and Psychological Measurement*, 37, 1031-1038.

- Stillman, P., Burpeau-Di Gregorio, M., Nicholson, G., Sabers, D., & Stillman, A. (1983). Six years of experience using patient instructors to teach interviewing skills. *Journal of Medical Education*, 58(December), 941-945.
- Stillman, P., Regan, M. B., & Swanson, D. (1987). Impact of several variables on physical examination skills of medical students. *Journal of Medical Education*, 62(November), 937-939.
- Stillman, P., Regan, M. B., Swanson, D., Case, S., McCahan, J., & al., e. (1990). An assessment of the clinical skills of fourth-year students at four New England medical schools. *Academic Medicine*, 65(5), 320-326.
- Stillman, P., Regan, M., Swanson, D., & Haley, H. (1992). Does gender differences in clinical skills as measured by an examination using standardised patients. In I. Hart, R. Harden, & J. Des Machasis (Eds.), *Current developments in assessing clinical competence* (pp. 390-395). Montreal: Can-Heal.
- Stillman, P., Ruggill, J., & Sabers, D. (1978). The use of practical instructors to evaluate a complete physical examination. *Evaluation and the Health Professions*, 1(1), 49-54.
- Stillman, P., Ruggill, J., Rutatla, P., & Sabers, D. (1980). Patient instructors as teachers and evaluators. *Journal of Medical Education*, 55(March), 186-193.
- Stillman, P., Swanson, D., Regan, M., Philbin, M., Nelson, V., & al., e. (1991). Assessment of clinical skills of residents utilizing standardised patients. *Annals of Internal Medicine*, 114(5), 393-401.
- Stillman, P., Swanson, D., Smee, S., Stillman, A., Ebert, T., & al., e. (1986). Assessing clinical skills of residents with standardised patients. *Annals of Internal Medicine*, 105, 762-771.
- Stratford, P., Thomson, M., Sanford, J., Saarinen, H., Dilworth, P., Solomon, P., Nixon, P., Fraser-MacDougall, V., & Pierce-Fenn, H. (1990). Effect of station examination item sampling on generalizability of student performance. *Physical Therapy*, 70(1), 31-36.
- Strauss, A., & Corbin, J. (1990). *Basics of Qualitative Research: Grounded Theory Procedures an Techniques*. Newbury Park, California: Sage.
- Strohm-Kitchener, K. (1983). Cognition, metacognition and epistemic cognition. *Human Development*, 26, 222-232.
- Sullivan, E. (1996). Teaching financial statement analysis: a cooperative learning approach. *Journal of Accounting Education*, 14(1), 107-111.
- Sullivan, H. (1953). *The Interpersonal Theory of Psychiatry*. (Rev. Ed. ed.). New York: Norton.
- Swanson, D., & Norcini, J. (1989). Factors influencing reproducibility of tests using standardised patients. *Teaching and learning in medicine*, 1(3), 158-166.
- Swanson, D., & Stillman, P. (1990). Use of standardised patients for teaching and assessing clinical skills. *Evaluation and the Health Professions*, 13(1), 79-103.
- Swanson, D., Norman, G., & Linn, R. (1995). Performance based assessment: lessons from the health professions. *Educational Researcher*, 24(5), 5-11, 35.

- Swanson, H. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of Educational Psychology*, 82(2), 306-314.
- Swartz, M., & Colliver, J. (1996). Using standardised patients for assessing clinical competence. *The Mount Sinai Journal of Medicine*, 63(3&4), 241-249.
- Swartz, M., Colliver, J., Cohen, D., & Barrows, H. (1994). *The effect of deliberate, excessive violations of test security on a standardised-patient examination: an extended analysis*. Paper presented at the Sixth Ottawa Conference on Medical Education, Toronto, Canada.
- Tamblyn, R., Abrahamowicz, M., Berkson, L., Dauphinee, D., Gayton, D., Grad, R., Isaac, L., Marrache, M., McLeod, P., & Snell, L. (1992). First visit bias in the measurement of clinical competence with standardised patients. *Academic Medicine*, 67(10 Supplement), S22-S24.
- Tamblyn, R., Klass, D., Schnabl, G., & Kopelow, M. (1991). The accuracy of standardized patient presentation. *Medical Education*, 25, 100-109.
- Tamblyn, R., Klass, D., Schnabl, G., & Kopelow, M. (1991b). Sources of Unreliability and Bias in Standardized-Patient Rating. *Teaching and Learning in Medicine*, 3(2), 74-85.
- Tamblyn, R., Schnabl, G., Klass, D., Kopelow, M., & Marcy, M. (1988). *How Standardized are Standardized Patients?* Paper presented at the 27th Research in Medical Education Conference, Washington, DC.
- Terry, W., & Higgs, J. (1993). Educational Programmes to Develop Clinical Reasoning Skills. *Australian Physiotherapy Journal*, 39(1), 47-51.
- Thomas-Edding, D. (1987). *Clinical problem solving in physical therapy and its implications for curriculum development*. Paper presented at the Tenth International Congress of the World Confederation for Physical Therapy, Sydney, Australia.
- Thomas-Edding, D. (1987). *Problem-solving in physical therapy: Implications for curriculum development*. Unpublished Ph.D., University of Toronto.
- Tichenor, C., Davidson, J., & Jensen, G. (1995). Cases as Shared Inquiry: Model for Clinical Reasoning. *Journal of Physical Therapy Education*, 9(2), 57-62.
- Tinning, R., & Fitzclarence, L. (1982). Students supervising students: An alternative approach to practicum supervision. *Australian Journal of Teaching Practice*, 2(2), 91-100.
- Topping, K. (1992). Cooperative learning and peer tutoring: An overview. *The Psychologist*, 5(April), 151-161.
- Topping, K. (1996). The effectiveness of peer tutoring in further and higher education: a typology and review of the literature. *Higher Education*, 32, 321-345.
- Topping, K., & Ehly, S. (1998). Introduction to peer assisted learning. In K. Topping & S. Ehly (Eds.), *Peer Assisted Learning* (pp. 1-23). London: Lawrence Erlbaum and Associates.

- Topping, K., Hill, S., McKaig, A., Rogers, C., Rushi, N., & Young, D. (1997). Paired reciprocal peer tutoring in undergraduate economics. *Innovations in Education and Training International*, 34(2), 96-112.
- Traband, M., & Dunn, T. (1988). Differences in clinical simulation performance: a role for advice-strategies? *Respiratory Care*, 33(9), 779-785.
- Traub, R., & Rowley, G. (1991). Understanding Reliability. *Educational Measurement: Issues and Practice*, 10(Spring), 37-45.
- Trautwein, S., Racke, A., & Hillman, B. (1996). Cooperative learning in the anatomy library. *Journal of College Science Teaching*, 26(3), 183-188.
- Triggs-Nemshick, M., & Shepard, K. (1996). Physical therapy clinical education in a 2:1 student-instructor education model. *Physical Therapy*, 76(9), 968-981.
- Umphred, D. (1995). Physical therapy differential diagnosis in the clinical setting. *Journal of Physical Therapy Education*, 9(2), 39-45.
- Vaidya, S. (1994). Improving teaching and learning through peer coaching. *Education*, 115(2), 241-245.
- van der Vleuten, C., & Newble, D. (1995). How can we test clinical reasoning? *The Lancet*, 345(April, 22), 1032-1034.
- van der Vleuten, C., & Swanson, D. (1990). Assessment of Clinical Skills with Standardized Patients: State of the Art. *Teaching and Learning in Medicine*, 2(2), 58-76.
- van der Vleuten, C., Norman, G., & De Graaf, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education*, 25, 119-126.
- Van Luijk, S., & Van der Vleuten, C. (1990). *A comparison of checklists and rating scales in performance-based testing*. Paper presented at the Current Developments in Assessing Clinical Competence, Ottawa, Canada.
- Van Rossum, H., Briet, E., Bender, W., & Meinders, A. (1990). *The transfer effect of one single patient demonstration on diagnostic judgement of medical students: both better and worse*. Paper presented at the Teaching and Assessing Clinical Competence, Groningen, The Netherlands.
- Vernon, D., & Blake, R. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine*, 68(7), 550-563.
- Vockell, E., & Asher, J. (1995). *Educational Research*. (2nd ed.). Englewood Cliffs, New Jersey: Prentice-Hall Inc.
- Vu, N. V., Steward, D., & Marcy, M. (1987). An assessment of the consistency and accuracy of standardized patients' simulations. *Journal of Medical Education*, 62(December), 1000-1002.
- Vu, N., & Barrows, H. (1994). Use of standardised patients in clinical assessments: Recent developments and measurement findings. *Educational Researcher*, 23(3), 23-30.
- Vu, N., Marcy, M., Colliver, J., Verhulst, S., Travis, T., & Barrows, H. (1992). Standardized (simulated) Patients' Accuracy in Recording Clinical Performed Check-List Items. *Medical Education*, 26, 99-104.

- Vygotsky, L. (1978). *Mind in Society*. Cambridge, MA.: M.I.T. Press.
- Vygotsky, L. (1986). *Thought and Language*. Cambridge, MA.: M.I.T. Press.
- Wagstaff, M. (1989). The team concept in supervised practice: benefits for students and preceptors. *Journal of the American Dietetic Association.*, 89(1), 78-81.
- Watson, S., & Marshall, J. (1995). Effects of cooperative incentives and heterogenous arrangement on achievement and interaction of cooperative learning groups in a college life science course. *Journal of Research in Science Teaching*, 32(3), 291-299.
- Watts, N. (1985). Decision analysis: a tool for improving physical therapy practice and education. In S. Wolf (Ed.), *Clinical Decision Making in Physical Therapy* (pp. 7-23). Philadelphia: F.A. Davis Company.
- Webb, N. (1982). Student interaction and learning in small groups. *Review of Educational Research*, 52(3), 421-455.
- Whelan, G. (1988). Improving medical students' clinical problem-solving. In P. Ramsden (Ed.), *Improving Learning: New Perspectives* (pp. 199-214). London: Kogan Page.
- Williams, L. (1995). Modified teaching clinic: peer group supervision in clinical training and professional development. *American Journal of Speech-Language Pathology*, 4(3), 29-38.
- Williams, R., Barrows, H., Vu, N., Verhulst, J., Colliver, J., Marcy, M., & Steward, D. (1987). Direct, standardized assessment of clinical competence. *Medical Education*, 21, 482-489.
- Williams, R., Lloyd, J., & Simonton, D. (1990). *Sources of OSCE examination information and perceived helpfulness: a study of the grapevine*. Paper presented at the Current Developments in Assessing Clinical Competence, Ottawa, Canada.
- Williamson, L., & Russell, D. (1990). Peer coaching as a follow-up to training. *Journal of Staff Development*, 11(2).
- Wilson, J., Stull, W., & Vinsonhaler, J. (1996). Rethinking cooperative education. *Journal of Cooperative Education*, 31(2), 154-165.
- Wood, D., & O'Malley, C. (1996). Collaborative learning between peers. *Educational Psychology in Practice*, 11(4), 4-9.
- Woodward, C., McConvey, G., Neufeld, V., Norman, G., & Walsh, A. (1985). Measurement of Physician Performance by Standardized Patients. *Medical Care*, 23(8), 1019-1027.
- Woolliscroft, J., Swanson, D., & Case, S. (1992). *Validity of extended matching and short answer response formats with pattern recognition items*. Paper presented at the Approaches to the Assessment of Clinical Competence, Dundee, Scotland.
- Worrell, P. (1990). Metacognition: implications for instruction in nursing education. *Journal of Nursing Education*, 29(4), 170-175.

- Wynn, M., & Kromrey, J. (1999). Paired peer placement with peer coaching: in early field experience: results of a four-year study. *Teacher Education Quarterly*, 26(1), 21-38.
- Yerks, R., & Dodson, J. (1908). The relation of strength of stimulus to rapidity of habit formation. *Journal of Comparative Neurology and Psychology*, 18, 459-482.
- Yinger, R. (1986). Examining thought in action: a theoretical and methodological critique of research on interactive teaching. *Teacher and Teacher Education*, 2, 263-282.
- Ytterberg, S., Harris, I., Allen, S., Anderson, D., Kofron, P., Kvasnicka, J., McCord, J., & Moller, J. (1998). Clinical confidence and skills of medical students: use of an OSCE to enhance confidence in clinical skills. *Academic Medicine*, 73(10).

Appendix 1: History Checklist

Student(s) Name/Code: _____

Duration of Encounter: _____ Date: _____

History Checklist: Shoulder Pain Case

0 = not done 1 = done poorly 2 = done correctly

0	1	2	Right or left handed
0	1	2	Asks you to describe the location of pain
0	1	2	Asks if you have pain anywhere else
0	1	2	Characteristics and nature of the pain (eg. sharp/dull, constant/periodic, superficial/deep)
0	1	2	**Asks you to explain how your pain developed
0	1	2	**Works through how your pain progressed initially to present status
0	1	2	Gets you to rate the severity of your pain on a 1-10 scale
0	1	2	**Inquires if you have pins or needles or numbness
0	1	2	Asks if you are taking any medications <u>and</u> what effect they have
0	1	2	**Inquires about the nature of any night pain
0	1	2	Inquires about the nature of any day pain
0	1	2	**Asks whether you have morning stiffness
0	1	2	Asks you to describe activities/postures that aggravate your pain
0	1	2	**Asks you how quickly your pain settles after causing it to occur (irritability)
0	1	2	Asks you to describe activities/postures that ease your pain
0	1	2	**Inquires about your past medical history: shoulder & other conditions (eg. other injuries, surgeries)
0	1	2	**Inquires about your general health <u>and</u> if any medical problems
0	1	2	Unexplained weight loss
0	1	2	Asks/determines whether you have had shoulder X-rays and the result of the x-ray
0	1	2	Asks/determines if you have had any other investigations (i.e., ultrasound test) and the results of this test.
0	1	2	Asks you what is your occupation
0	1	2	Asks for specific details about the nature your job: eg. physical demands
0	1	2	Asks you whether you play any sport and the frequency of this activity.
0	1	2	Inquires about your social situation

Total Score _____ /48

History Checklist Guidelines for Simulated Patient

- **Asks you to describe your main complaint:** asks you to describe your problem in your own words
- **Asks you whether you are right or left handed**
- **Asks you to describe the location of pain:** asks you to describe exactly where the pain is located around the shoulder area
- **Asks if you have pain anywhere else:** asks if you have pain any where other than the shoulder region - eg. alongside the neck, down the arm, the elbow, across the back or chest.
- **Characteristics and nature of the pain:** asks you whether the pain is deep/superficial, dull/sharp, constant/periodic. Basically, they try to get you to describe the quality of your pain. Should ask at least 2 of these categories for a score of two.
 - Zero points if they just ascertain that you were hammering.
 - One point if they ascertain that it was upward hammering.
 - Two points if they ascertain that it was prolonged hammering.
- **Works through how your pain progressed initially to present status:** the therapist should get you to describe how your pain has changed since the initial episode of hammering through to how it feels at present. They should ask whether your pain is getting better or worse. In your case, the pain developed two days later as an ache and quickly became very intense. Since seeing the Doctor, the pain has subsided.
- **Gets you to rate the severity of your pain on a 1-10 scale:** the therapist asks you to rate your pain on a scale of 1 - 10. They should attempt to get you to somehow describe the intensity of your pain relative to having no pain.
- **Inquires if you have pins, needles or numbness:** therapist should ask you if you are having/had these sensations in your left arm. Two points for all 3 items.
- **Asks if you are taking any medications and what effect they have:** the therapist should ask you both components of this question. Otherwise score 1 if they only ask the first part of the question.
- **Inquires about the nature of any night pain:** the therapist should ask you if you have pain currently at night (score 1 if the inquiry ends here). They should also explore this night pain a bit. For example, does the night pain itself wake you up or is it related to movement or laying on the left shoulder? (score 2 points)
- **Inquires about the nature of any day pain:** the therapist should ask you how your shoulder feels as the day progresses. They should probe to see what causes your pain to increase by the end of the day for a full score of 2 points.

- **Asks whether you have morning stiffness:** the therapist should ask you whether your shoulder is stiff when you wake up in the morning. They only score 1 point if they do not inquire about current morning stiffness.
- **Asks you to describe activities/postures that aggravate your pain:** the therapist should ask you what movements and activities cause you to have the pain in your shoulder. For 2 points they should encourage you to identify at least one functional activity.
- **Asks you how quickly your pain settles after causing it to occur (irritability):** After doing something that causes shoulder pain, the therapist should inquire how quickly this pain settles (1 point only). The therapist should also inquire about repeated activities that cause pain and how quickly the pain settles. Both components should be asked for a score of 2 points.
- **Asks you to describe activities/postures that ease your pain:** the therapist should ask you if there is anything that eases your pain.
- **Inquires about your past medical history:** the therapist should ask you if you have ever had shoulder troubles (score 1 point only if they stop here). For two points they should ask if you have had any other injuries/problems with your other bones/joints and should ask you if you have had any previous surgeries.
- **Inquires about your general health and if any medical problems:** the therapist should ask you if you are healthy (score 1 point if stop here) and whether you have any other medical conditions - eg. high blood pressure, heart problems, diabetes, breathing problems.
- **Unexplained weight loss:** should ask whether you have lost any weight recently that was unplanned
- **Asks/determines whether you have had shoulder X-rays and the result of the x-ray:** Both components of this statement must be asked for a score of 2.
- **Asks/determines whether you have had any other investigations (eg. ultrasound test) and the results of this test:** for a score of 2 points, the therapist should ask if you have had any other tests for the shoulder and the results.
- **Asks you what is your occupation:** the therapist should ask you what you do for a living
- **Asks for specific details about the nature of your job: eg. physical demands:** the therapist should ask you to describe the sorts of things you would do on a daily basis in your job.
- **Asks you whether you play any sport and the frequency of this activity:** the therapist should ask you if you play any sport (score 1 point if this is all they ask). They should also ask whether you have played golf recently and whether your shoulder ever gave you trouble playing this sport for 2 points.
- **Inquires about your social situation:** the therapist should find out a bit about your home situation (eg. you own a home in Perth and share with a friend, work 3 weeks on 1 week off, that you are concerned about getting better quickly).

Appendix 2: Physical Examination Checklist

Student(s) Name/Code: _____

Duration of Encounter: _____ Date: _____

Physical Examination Checklist: Shoulder Pain Case

0 = not done 1 = done poorly 2 = done correctly

0	1	2	Asks you to remove your shirt
0	1	2	Observes posture from front and back
0	1	2	Asks you to reach behind your neck and behind your back with your left and right arm.
0	1	2	**Gets <u>you</u> to actively perform shoulder movements in either sitting or standing
0	1	2	** <u>Therapist</u> passively moves your shoulder through the various movements while you are laying down
0	1	2	**Tests isometric strength of your shoulder muscles in sitting or standing
0	1	2	Tests the isometric strength of the biceps muscle that bends your forearm
0	1	2	Therapist gets you to pull your shoulder blades together in sitting/standing
0	1	2	Therapist gets you to pull your shoulder blades together laying face down and applies some resistance.
0	1	2	Therapist gets you to raise the point of your elbow or arm towards the ceiling and applies some resistance
0	1	2	**Checks your neck movements
0	1	2	**Presses on the outer point of your shoulder (pain no. 1)
0	1	2	**Presses on the front of the shoulder (pain no. 2)
0	1	2	Feels on the boney prominence on the top of your shoulder
0	1	2	**Checks for tenderness by feeling the muscles above your shoulder blade (pain no. 3)
0	1	2	**Tests whether you can hold your arm up in the air, then determines how this changes when thumb is turned downward (empty can position)
0	1	2	<u>Therapist</u> holds your arm on theirs and then passively twists it in a downwards motion.
0	1	2	**Tests the Pain No. 2 area using 4 possible tests
0	1	2	** <u>Therapist</u> manipulates your shoulder while you are laying on your back
0	1	2	** <u>Therapist</u> stresses the joint at the top of your shoulder by pressing on the pointy bits of bone on top of your shoulder

Total Score _____ /40

Physical Exam. Checklist Guidelines for Simulated Patient

- **Asks you to remove your shirt:** shirt should be removed for examination for a score of 2.
- **Observes posture from front and back:** the therapist should get you to stand up in a relaxed posture and look at your posture from the front and the back. They may or may not actually state they are observing your posture.
- **Asks you to reach behind your neck and behind your back with your left and right arm:** - self-explanatory
- **Gets you to actively perform shoulder movements in either sitting/standing:** movement forward, sideways and rotation in/out should be tested and the therapist must also ask you to report any pain that you have during these movements for a full score of 2 points.
- **Therapist passively moves your shoulder through the various movements while you are laying down:** the therapist should move your arm forward, sideways, and rotate it inwards and outwards and they must also ask you to report any pain during any of the movements for a full score of 2 points.
- **Tests isometric strength of your shoulder muscles in sitting or standing:** the therapist should check the isometric strength of forward, sideways, and turning in and out movements and should ask you to report if you have any pain during these movements for 2 points.
- **Tests the isometric strength of the biceps muscle that bends your forearm:** the therapist should check the isometric strength of your biceps muscle by getting you push up against resistance with a bent elbow. The therapist should ask you to report if you have any pain during these movements for a score of 2.
- **Therapist gets you to pull your shoulder blades together in sitting/standing:** the therapist should ask you pull your shoulder blades together. They may do this one at a time or both together. They may also do this test by having you extend your arm forward while they check your shoulder blade position. Whichever tests they choose, they must evaluate both shoulder blades for a score of 2 points.
- **Therapist gets you to pull your shoulder blades together laying face down and applies some resistance:** the therapist will have you lay on your stomach with your arms out to the side as best you can and will ask you to pull your shoulder blades together. They may do this one at a time or both together. They should apply some resistance (at the arm or at the shoulder) as well. They must check both sides for a score of 2 points and must apply some resistance for a full score.
- **Therapist gets you to raise the point of your elbow or arm towards the ceiling and applies some resistance:** the therapist will have you lay on your back with your elbow pointing towards the ceiling with your elbow fully bent or with your arm straight, reaching towards the ceiling. They should get you to push your upper limb upwards while applying resistance for a score of 2 points.

- **Checks your neck movements:** the therapist should get you to move your head in all directions through the full range (forward, backward, left turn, right turn, left sidebend and right sidebend). While they may not add overpressure at the end of each movement, for a score of two points they must at least add some overpressure to your neck when you are side bending your head towards the right.
- **Presses on the outer point of your shoulder (pain no. 1):** the therapist should press adequately on the bone just below and in front of the shelf of bone on the top of your shoulder. The arm can be by your side or placed slightly backward.
- **Presses on the front of the shoulder (pain no. 2):** the therapist should press adequately on the front of shoulder directly above the biceps muscle.
- **Feels on the boney prominence on the top of your shoulder:** the therapist should press adequately on the boney prominence on the top of the shoulder.
- **Checks for tenderness by feeling the muscles above your shoulder blade (2 spots) : (pain no. 3):** the therapist should palpate the muscles just to the left of your neck, and to the left and slightly below the base of your neck for a score of 2 points.
- **Tests whether you can hold your arm up in the air, then determines how this changes when thumb is turned downward (empty can position) - For one point:** The arm is raised out to your side (but not into your painful position) while you are instructed to hold it there for a few moments while the therapist inquires about your pain. For the second point: You are then instructed to turn your thumb downward (as if emptying a can) and report on your pain. If these tests are done with your arm close to or actually within your painful position then the therapist only scores one point.
- **Therapist holds your arm on theirs and then passively twists it in a downwards motion:** the therapist must hold up the weight of your arm and get you to relax. They then must rotate your arm downwards and ask you how your shoulder feels for a score of two points.
- **Tests the Pain No. 2 area using 4 possible tests**
 - Presses on your pain no. 2 area while getting you to tighten your biceps muscle against resistance. The therapist at the same time will try to turn your upper limb outward
 - The therapist positions your straight arm backwards and tries to get you to bring your arm forward against resistance (adequate resistance is necessary for a score of 2)
 - The therapist positions your straight arm by your side with your thumb backward and gets you to bend your elbow through range against resistance (adequate resistance is necessary for a score of 2 points)
 - Your straight arm is brought forward with your palm facing upwards and resistance is applied to your arm (adequate resistance is necessary for a score of 2 points)

- **Therapist manipulates your shoulder while you are laying on your back:** for this manipulation to be done properly the therapist must press downward on your shoulder (from the top), with your arm positioned out to the side just before where you normally experience pain. This must be compared to the other side for a full score of 2 points.
- **Therapist stresses the joint at the top of your shoulder by pressing on the pointy bits of bone on top of your shoulder:** the therapist should apply stress to this joint by pressing on the bones at the top of your shoulder. This is different from just feeling the joint, they must apply some force. This test is usually done while you are laying on your back but may be done in other positions.

Appendix 3: Arizona Clinical Interview Rating Scale

Student Name/Code _____ Date _____

Score /36 Arizona Clinical Interviewing Rating Scale Page 1/2

1. Organisation

2	The interviewer(s) elicit(s) detailed documentation of the patient's current problem (onset, duration, location, quality, intensity, setting, significant positive, significant negatives, aggravating and alleviating factors) and all additional data relevant to the problem under discussion (past medical history, social history).
1	The interviewer(s) elicit(s) detailed documentation of the patient's current problem but fails to obtain all additional data relevant to the problem under discussion.
0	The interviewer(s) fail(s) to elicit detailed documentation of the patient's current problem.

2. Timeline

2	The interviewer(s) obtain(s) information pertaining to the chief complaint and history of the present illness in a logical, systematic and orderly progression, gathering all necessary information, starting with the first signs and symptoms of current illness and following their progression to the present and obtains most of the pertinent information.
1	At times, the interviewer(s) do(es) not obtain information pertaining to chief complaint and history of presenting complaint in chronological order, but is still able to obtain most of the pertinent information.
0	The interviewer(s) obtain(s) information pertaining to the chief complain and history of present illness in a haphazard and unrelated fashion, resulting in the omission of pertinent data.

3. Questioning Skills

2	The student(s) start(s) information gathering with open ended questions and uses direct and forced choice questions when he/she needs to narrow in on the pertinent positive and negative points that need further elaboration.
1	The interviewer(s) often fail to begin a line of inquiry with open-ended questions but rather employs direct and forced-choice questions to obtain information .
0	The interviewer asks many leading questions, why questions and multiple questions, for example: you have never had pain before, have you?, that pain wasn't always that bad, was it?

4. Questioning Skills

2	The interviewer(s) ask(s) questions and/or takes notes in a manner which results in an interview that progresses smoothly with few unnecessary delays in the dialogue.
1	At times, the interview is marked with unnecessary pauses which temporarily break the continuity of the interview.
0	The interview is conducted in such a manner that long pauses occur which break the continuity of the interview.

5. Questioning Skills

2	The interviewer(s) do(es) not repeat questions, seeking duplication of information that has previously been provided, unless clarification or summarisation of previous information is necessary.
1	The interviewer(s) occasionally repeats questions seeking the duplication of information. Such information is not sought for the summarisation or clarification of information, but is the result of the interviewers failure to remember the data.
0	The interviewer frequently repeats several questions, seeking information previously provided because he/she fails to remember data already obtained.

6. Questioning Skills

2	Questions asked, as well as information provided to the patient during the interview, are concise and easily understandable; content is free of difficult medical terms and jargon. If jargon is used, the words are immediately defined for the patient.
1	The interviewer(s) sometimes uses medical jargon during the interview, failing spontaneously to define the medical terms for the patient unless specifically requested to do so by the patient.
0	Questions asked, as well as information provided to the patient during the interview, are confusing and difficult to understand; content contains numerous difficult medical terms and jargon.

7. Documentation of Data

2	The interviewer(s) seek(s) specificity and verification of the patient's responses.
1	The interviewer(s), at times, seek specificity and verification of the patient's responses, but not always.
0	The interviewer(s) does not verify the patient's responses, accepting information at face value.

8. Rapport

2	The interviewer(s) maintain(s) good eye contact with the patient during the interview.
1	The interviewer(s) maintain(s) some eye contact: however, the frequency could be increased.
0	The interviewer(s) make(s) no attempt to maintain eye contact with the patient.

9. Rapport

2	The interviewer(s) is/are attentive to the responses of the patient and allows her to complete statements without undue interruptions.
1	The interviewer(s) is/are usually attentive but on a few occasions interrupts the patient unnecessarily.
0	The interviewers) is/are nothing more than a detached data-gatherer, frequently interrupting the patient's statements without allowing her to complete her train of thought.

10. Rapport

2	The interviewer(s) seem(s) alert, sensitive and responsive to possible concerns expressed by the patient regardless of whether such concerns are immediately relevant to the patient's present physical problem (i.e., domestic problems, depression, psychosocial issues) and is able to explore them in sufficient depth
1	The interviewer(s) is/are able to detect concerns expressed by the patient unrelated to the present physical problems, but fails to explore them in sufficient depth.
0	The interviewer(s) seem(s) unalert and/or insensitive to possible concerns expressed by the patient if such concerns are not directly related to the present physical problem. For whatever reasons, the interviewer(s) tend(s) to avoid discussing possible problem areas which could have either immediate or future implications for the mental or physical health of their patient.

11. Rapport

2	The interviewer(s) provide the patient with intermittent positive social reinforcement and feedback (i.e., occasional smile, nodding the head in a positive manner, praising the patient for appropriate technique)
1	The interviewer(s) is/are neither overly positive nor negative in dispensing feedback and reinforcement
0	The interviewer(s) provide(s) the patient with little support or positive social reinforcement. Stress is on negative rather than positive attributes of the patient.

12. Rapport

2	At the end of the complete interview, the interviewer(s) encourage(s) the patient to discuss any additional point or ask additional questions and provides the patient with an adequate opportunity to do so.
1	At the end of the complete interview, the interviewer(s) provide the patient with the opportunity to discuss any additional points or ask additional questions but neither encourages or discourages her.
0	At the end of the complete interview, the interviewer(s) do(es) not provide the patient the opportunity to discuss any additional points or ask any questions.

Stillman P. (1980) Arizona Clinical Interview Medical Rating Scale, *Medical Teacher*; 8-31.

Appendix 4: Post-Encounter Probe

Post-Encounter Probe: Questions and Solutions

Diagnosis (maximum 8)

- Impingement Syndrome (subacromial) or 2. Supraspinatus Tendinitis or 3. Calcific Tendinitis or 4. Incomplete Tear of the Rotator Cuff or Rotator Cuff Tear (4)
- Bicipital Tendinitis (4)

Differential Diagnoses (maximum 10 - 2 each)

- + Cervico-Thoracic Spinal Stiffness (T2)
- + Tightness/Spasm of the Levator Scapulae and Supraspinatus Muscle Bellies
- + Transverse Ligament Sprain
- + Dysfunctional scapulo-thoracic mechanics
- +Subacromial/Subdeltoid Bursitis

Diagnostic Errors (-1 mark for each of these answers)

- - Sub-coracoid impingement syndrome
- - Acromio-clavicular strain
- - Capsulitis
- - Glenohumeral Instability - peri-articular laxity
- - Glenohumeral Intra-articular pathology - labral tear
- - Degenerative Arthritis
- - Biceps tendon tear
- - Cervical rib
- - Pectoralis minor syndrome
- - Claviculo-costal syndrome
- - Somatic referral
- - Cervical nerve root impingement

Management (maximum 20)

- (2) Accessory movements/mobilisation of the glenohumeral joint: (AP, PA or longitudinal/caudad acceptable answers)
- (2) Electrotherapeutic Modality: pulsed ultrasound, laser, interferential or high voltage galvanic
- (2) Scapulothoracic strengthening/stabilisation exercises: - middle/lower trapezius/serratus anterior - should precede resisted exercises
- (2) Resisted exercises - internal and external rotation: theraband, may need to start with isometric progressing to isotonic; Allingham program
- (1) Biceps strengthening exercises
- (2) Mobilisation: T2
- (2) Soft Tissue Massage: Levator Scapulae/Supraspinatus
- (2) Stretching tight muscle groups: Levator Scapulae/Pectoralis Minor/Upper Trapezius
- (2) Postural re-education: stretching tight muscles and strengthening weak muscles
- (3) Home Program: home exercises - resisted using theraband: use of ICE if pain increases: advice re: avoiding painful movements and aggravating activities

Post-Encounter Probe Solutions

(bold denotes key feature = 2)

(normal text denotes important feature but not key = 1)

(italics denotes distracter = 0)

4. Select 5 key history findings from the list that lead to the patient's correct diagnosis(es)/problem(s).

- *Able to lay on either right or left side while sleeping*
- Full and pain free cervical range of motion
- **Non-steroidal anti-inflammatory medications help ease left shoulder pain**
- **Pain developed after prolonged upward hammering**
- Current pain is superficial and sharp
- *There is no morning stiffness*
- *Pain in the left shoulder still wakes the patient up at night*
- **Reaching behind back with left arm is painful**
- Sporting activity potential factor causing current problem
- *Occasional clicking noise in left shoulder with movements in and out of elevation*

- **Moving the left arm forward is painful**
- Pain increases in left shoulder as day progresses
- *Working with hands in fixed spinal/arm positions elicits shoulder ache*
- No history of left shoulder pain
- **Significant ultrasound findings**

5. Select 5 key history findings from the list that lead to the patient's correct diagnosis(es)/problem(s).

- Pain in left shoulder is improving since initial onset
- Patient is right handed
- **Significant x-ray findings**
- Pain developed a couple of days later in the left shoulder
- *Repeated movements of the left shoulder cause considerable shoulder pain and it takes several hours for the pain to subside*
- **Moving the left arm out sideways is painful**
- *Pins and needles sensation occasionally felt on left outer arm and forearm*
- **Pain does not extend below deltoid insertion (left side)**
- *Previous history of neck problems*
- There are no significant medical problems
- *The pain is constant*
- **Unable to sleep on affected side**
- Varying intensity of pain ranging from 0 - 5 on a 10 point scale
- *X-rays normal*
- **Occupational activities a causal factor**

6. Select 5 key physical examination findings from the list that are central in leading to the patient's correct diagnosis(es)/problem(s).

- *Grade 3+ central P-A accessory movement testing of Cervical 5/6 yields stiffness/discomfort*
- **Painful passive left shoulder flexion**
- *Difficulty initiating left shoulder abduction*
- Normal biceps tendon reflexes
- **Exquisite tenderness over the left greater tuberosity**

- Full Cervical Spine Range of Motion
- *Left shoulder: limitation of movement in capsular pattern*
- Mild tenderness over the left greater tuberosity
- **Tenderness over the left upper trapezius muscle belly**
- No pain or tenderness at acromioclavicular joint
- *Cervical Left Side Flexion Limited*
- *Strong and pain free isometric shoulder flexion (Speed's Test)*
- **Positive Hawkins-Kennedy/Impingement Test (passive shoulder flexion to 90 degrees combined with elbow flexion to 90 degrees, followed by internal rotation of glenohumeral joint)**
- No pain or resistance on grade 3 A-P left glenohumeral accessory movement testing
- **Weak and painful isometric left shoulder abduction**

7. Select 5 key physical examination findings from the list that are central in leading to the patient's correct diagnosis(es)/problem(s).

- No tenderness on palpation over the lesser tuberosity of the humerus
- Painful passive shoulder abduction
- **Slightly painful resisted elbow flexion with palpation over bicipital groove (Yergason's Test)**
- *Serratus anterior muscle strength grade 5 bilaterally*
- **Painful arc**
- Cervical right side flexion with overpressure produces discomfort along the left side of the neck/nape of the neck
- *Left shoulder level slightly lower than right level*
- *Full active external rotation left shoulder*
- **Positive Supraspinatus Test (resisted abduction/neutral rotation and/or resisted abduction/internal rotation)**
- *Painful but strong isometric internal rotation of the left shoulder*
- No pain, resistance or stiffness on Cervico-Thoracic Accessory Movement testing (central P-A movements)
- **Altered left sided scapulothoracic rhythm**
- *Weak and painful isometric left shoulder adduction*
- **Grade 3 caudad left glenohumeral accessory movement testing reduces/lessens pain**

- Painful passive external rotation

8. Select 5 key physical examination findings from the list that are central in leading to the patient's correct diagnosis(es)/problem(s).

- *Strong and pain free isometric left external rotation*
- **Painful and limited active shoulder flexion**
- *Positive apprehension test: left shoulder*
- *Painful but strong isometric shoulder adduction: left side*
- **Pec Minor Tightness**
- **Painful and limited active left shoulder abduction**
- Negative drop arm test
- Weak elbow flexors (left side)
- **Tenderness on palpation over the left bicipital groove**
- Postural changes
- Grade 4 strength left middle/lower trapezius
- **Strong, slightly painful isometric shoulder flexion (Speed's Test)**
- *No pain, resistance or stiffness on Cervico-Thoracic Accessory Movement testing (unilateral P-A movements): left and right side*
- Grade 3+ left unilateral P-A accessory movement testing of T2 yields stiffness/discomfort
- *Limited Cervical Right Rotation*

Appendix 5: Simulated Patient Training Notes

Simulated Patient Training Notes

Simulation Title: Supraspinatus Tendinitis, Left Shoulder

Setting: Outpatient Department

Age Range: 25 - 40

Physical Characteristics: male, average build, right handed

Presenting Complaint: Pain in left shoulder.

1.0 Present History:

Instructions for Responding to Therapist:

- *you volunteer information about your shoulder pain - i.e., location*
- *the remaining information must be elicited by questioning. To help you respond, information is presented as a series of bullets. Each bullet more or less represents an answer to a single question that the therapist may ask.*

1.1 Pain

- Located only along the outer shoulder. Starts at the shoulder and extends down no further than the upper 1/3 of the outer arm.
- Current pain is superficial, intermittent (i.e., only when you move it for most part); located under the acromion
- Current pain is sharp, with movement
- Current pain occurs only when you move your arm above the level of your shoulder. Reaching behind your lower back to scratch is uncomfortable and will reproduce your pain
- **You have no pain anywhere else, or any other symptoms (eg. ache, stiffness, clicks, clunks, pins and needles, numbness, sympathetic signs)**

1.2 Pain Rating

- *If you are asked to rate the pain:* there is no pain at the moment while the arm is resting. When you move your shoulder, the pain is around a 6 out of 10.
- When you first started to get the pain, it was around 8 on a scale of 10.

1.3 Onset and Progress of Shoulder Pain:

- Hammering in large bolts with a sledgehammer using an upward motion above your head (right arm closest to hammer head in more extension/neutral with left arm bent at elbow in more internal rotation at shoulder). This was about 10 days ago.
- The hammering activity occurred for about 30 minutes.
- You were still able to work after that for a couple of days without any problems as this hammering activity was an exception from your normal work activities. Most of your work is lighter than that.
- Two days after hammering, general muscle pain in the shoulder developed. A couple of days after that, **very** painful in the shoulder and was constant.
- Pain has subsided substantially over the past week and now only occurs when you move the shoulder.
- If asked how much your shoulder has improved since the beginning of the pain, you say it is about 60 per cent better.

1.4 Visit to your Doctor:

- You saw your Doctor here in Perth 4 days after the hammering incident (as you flew out for your normal time off 3 days after the hammering episode - you are on a 3 week on, 1 week off schedule).
- Your Doctor gave you a prescription for Feldene (anti-inflammatory) on day 4.
- The tablets relieved your shoulder pain considerably after a couple of days.
- You go back to see your Doctor in a couple of days.

1.5 Night pain:

- The pain in your shoulder did wake you up at night the first few days when you started to get the intense pain.
- It woke you up constantly (~4 times a night) and you were taking (panadeine forte) every 3 - 4 hours to relieve the pain.
- You are still unable to sleep on the left side due to pain. You sleep on your back or right side with one pillow under your head.
- Right now, can sleep through the night provided you do not lay on your left side.
- You have stopped taking the panadeine forte.

1.6 Day pain:

- The shoulder pain increases progressively throughout the day and is generally worse by the evening. Much depends upon how much you use the shoulder.

1.7 Morning stiffness:

- Your shoulder is not particularly stiff in the morning. Initially when the pain started it was a bit stiff in the morning but that disappeared in a couple of days.

1.8 Activities that aggravate the pain:

- Most movements that require you to move your arm above shoulder level are painful. The sideways movement causes the most pain, followed by the forward movement.
- Most activities of daily living that require you to use your left arm are painful - eg. lifting a heavy kettle or grocery bag if not careful, reaching behind your back, overhead activities. You can manage driving your car because of the power steering.
- When you do something that causes pain in the shoulder, the pain comes on immediately with the activity. When you stop doing it, the pain settles down quickly - about 10 seconds. Even if you were repeat the movement, the pain settles quickly although the amount of pain increases slightly with repeat movement.

1.9 Things that ease the pain:

- resting the arm on your lap, keeping it by your side, panadeine forte, feldene

2.0 Past History:

- You have never had this sort of pain in the shoulder before. You have never had a shoulder problem.

Simulated Patient Training Notes

3.0 Miscellaneous Information:

- General health - good, no heart or breathing problems, no diabetes, abdominal, visceral trouble. No previous cervical or thoracic problems.
- No pins or needles or sensory changes in arm
- No headaches
- Coughing/Sneezing: do not cause any pain
- No unplanned or unexplained weight loss
- No other medical or surgical problems - past or present, no other medications
- If the therapist asks you whether you have had x-rays you state that you have had X-rays of the Chest - 3 years ago - showed a broken 7th rib (right side - no separation) - from playing recreational football. Healed by itself.

- Therapist should ask you if you specifically had Xrays of the shoulder. These were ordered by Doctor: revealed calcium growth in the tendon
- Therapist should ask you if you specifically had any other investigations of the shoulder: you had an ultrasound of the shoulder: revealed calcium growth in the tendon.

4.0 Psychosocial:

4.1 Occupation:

- You are a tradesman (fitter), past 16 years
- Your job entails a lot of maintenance work such as pulling, pushing, lifting.
- You cannot work at the moment
- You are currently on compensation but are still awaiting your claim to be processed (it is now almost 2 weeks).
- You currently work on the mines (gold) just outside of Kalgoorlie. Surface work. You work three weeks on, and one week off.

4.2 Domestic:

- single, live with a flat-mate when in Perth.

4.3 Recreation:

- You play golf recreationally
- Right handed swing
- You haven't played any golf since hurting your shoulder. Last played 5 weeks ago during your time off.

5.0 Physical Examination

This section is part of the physical examination. You are only required to do what the therapist instructs you to do. Instructions for reproducing the correct signs and symptoms are described below. Instructions for some of the other tests and examination procedures will also be provided as part of the training.

5.1 Presentation:

5.2 Facial:

- You do not present as being in a lot of pain but you do wince when you have the pain in your shoulder

5.3 Body Language:

- Casual, slightly 'closed' body language

5.4 Posture:

- Normal posture, slightly rounded shoulders and left shoulder slightly elevated. Head slightly forward.

5.5 Undressing:

- If the therapist asks you to get undressed you have difficulty removing your shirt. Wear a buttoned shirt and take the right arm out first, then remove shirt off of the left arm with assistance of the right arm.

5.6 Cognitive/Affective:

- You are a bit stressed/frustrated about this injury as you are on a one year contract (currently in your fifth month).
- You are worried that once you return to work, they won't renew your contract because of concerns that you might get injured again. You want the therapist to guarantee that they can fix it completely and quickly.
- As far as you know, you have pulled a muscle and hope it will resolve quickly.
- You were hoping to work there for two to three years to improve your debt situation as you are paying off your mortgage for your own unit plus another unit that you have recently purchased to negatively gear your income.
- The compensation payments are still delayed and you are concerned about making some of your payments to creditors.

5.7 Objective Tests:

5.7.1 Movement of the Left Shoulder by Yourself:

(The pain is about 6/10 - when the arm is restored to the resting position you do not have any pain)

- forward: middle 1/3 of the shoulder movement is painful*, but then you can almost go the full range before pain again - coming back down the shoulder pain catches you again
- sideways: middle 1/3 of the shoulder movement is painful*, but then you can almost go the full range before pain again - coming back down the shoulder pain catches you again
- backwards: 95 % no pain
- horizontal across chest 85 % stopped by pain (minimal pain)
- arm in to stomach 100 % no pain
- arm out, away from stomach 40 % stopped by pain
- arm behind back you can only go as far as the middle of your back at the belt line.
- arm behind neck you can touch the big bone at the base of your neck but once your arm starts to go above the level of your shoulder you start to have pain and find it difficult to go further

* move your shoulder blade as a unit with your arm when performing these movements. If the therapist holds down the shoulder blade, you have less shoulder movement as a result

(Re: sideways movement - anytime you are put into the position with the arm rotated inward the pain is more intense)

5.7.2 Movement of the Left Shoulder by the Therapist:

(the pain in the shoulder that is experienced here is less in comparison to when you perform the movement yourself - about 5/10 - when the arm is restored to the resting position you do not have any pain)

- forward: middle 1/3 of the shoulder movement is painful*, but then you can almost go the full range before pain. If pressure is applied at the end you have a bit of pain and do not let the therapist go further. Coming back down the shoulder pain catches you again. PAIN NOT AS BAD WHEN THERAPIST DOES THIS MOVEMENT.
- sideways: middle 1/3 of the shoulder movement is painful*, but then you can almost go the full range before pain. If pressure is applied at the end you have a bit of pain and do not let the therapist go further. Coming back down the shoulder pain catches you again. PAIN NOT AS BAD WHEN THERAPIST DOES THIS MOVEMENT.
- backwards: 100% no pain, if pressure applied - it is okay
- across chest: (Horiz): 90 % stopped by pain (minimal pain), pressure by the therapist increases the discomfort slightly.
- arm turning in (by side) 100 % no pain
(90 degrees) pain in this starting position, with movement pain increases. 50 % maximum movement. Pressure through joint by therapist increases pain and you do not let the therapist move your arm further
- arm out (by side) 40 % stopped by pain
(90 degrees) pain in this starting position, with movement pain increases, 50% maximum movement. Pressure through joint by therapist increases pain and you do not let the therapist move your arm further

5.7.3 Strength of the Left Shoulder Blade Muscles

- Therapist may ask you to pull your shoulder back which you can do but a bit less intensely on the left
- Therapist may put you on your stomach and position your left arm as if you were reaching up for something. They will ask you to hold your arm against their resistance. You can do this but not as strongly as your right arm
- Therapist may put you on your back with your elbow pointing upwards and your elbow bent completely. They will ask you to raise your arm upward towards the ceiling against resistance. You can do this but not as strongly as your right arm.

5.7.4 Strength of the Left Shoulder Without Movement

(if strength is tested with movement you follow the same general pattern as below, except that the pain and movement follows the pattern indicated in 5.7.1)

- forward average but painful (slight-more over front of shoulder)
- sideways weak painful
- backwards strong no pain
- inwards strong no pain, although you can ‘feel’ your shoulder
- rotate out weak painful
- rotate in strong no pain, although you can ‘feel’ your shoulder

5.7.5 Movement and Strength of the Left Elbow

- You can move your elbow through its full range of motion. There is normal strength as well. The only thing of note is that when you are pushing your forearm up against resistance you have a slight amount of pain in the front of the shoulder.

5.7.6 Tightness of the Shoulder

- The therapist will lay you on your back with your arms by your side or slightly tucked under your bottom. They will note whether your left shoulder is higher than the right by measuring with their fingers or by pressing down and evaluating tightness. Your left shoulder should be slightly higher and if pressed upon, a bit uncomfortable just to the inside of your shoulder atop your chest.

5.7.7 Neck Movements:

- All movements of your neck are normal and painfree except when you side bend your head to the right. You feel just a bit of tension (not pain) along the left side of the neck (lower part)when bending in this direction. If the therapist adds some pressure to these movements they are still painfree. The left sided neck tension mentioned earlier, however, increases slightly when pressure is added by the therapist to right side bending.

6.0 Special Tests

6.1 Palpation:

- Tenderness when palpating the greater tuberosity.
 - Medium pain if the arm is at your side
 - Acute pain if the arm is positioned into extension (hand behind back psn.)
 - Minimal pain if the arm is in an abducted position.
 - If the arm is rotated while in any of the positions above, the intensity of the pain is the same when the greater tuberosity passes under the palpating finger(s).
- Tenderness when palpating the bicipital groove (Lippman' Test)
 - The amount of pain here is minimal but noticeable. Slight increase if the biceps is contracted and they rotate the arm while palpating.
- Tenderness bilaterally when palpating the mid-belly of levator scapulae
The left side, however, is slightly more tender
- Tender bilaterally when palpating the mid-belly of the supraspinatus
The left side, however, is slightly more tender
- **Palpation in other area is not tender (eg. infraspinatus, joint line, AC, sub-coracoid space)**

6.2 Special tests:

- Codman's test - patient is unable to hold their left arm up at the end range of their abduction due to pain if it is positioned at 90 degrees. Therapist must return it back to the patient's side.
- Supraspinatus test/Drop arm test/Empty can test - the patient's arm is abducted to 30 degrees. You can hold this position but you do have pain. You are unable to hold it up against any resistance. If the arm is internally rotated, the pain increases and the patient finds it difficult to hold the arm up independently. *(Anytime the arm is put into internal rotation, the pain is greater)*
- Hawkins and Kennedy Impingement test - the patient's arm is supported in the therapist's arm and the shoulder is pulled into internal rotation. You have pain when you reach the end of your movement but also have some pain by virtue of being at 90 degrees of abduction.
- Neer test - resisted horizontal flexion across chest - minor discomfort - (biceps tendon). Again, this position is painful because of the testing position
- Speed's test/Hawkins and Kennedy test - resisted shoulder flexion with elbow extension for biceps testing - reproduces some mild discomfort anterior shoulder. Also resisted elbow flexion, supination with palpation over bicipital groove reproduces some mild discomfort anterior shoulder (Yergason's).

6.3 Accessory Movement Tests: Shoulder (carried out just before P1, otherwise do not have any limitations or pain) P1 is in flexion or abduction

- P-A, AP, Caudad glenohumeral: anything less than a grade 3 does not produce pain. If the grade of mobilisation is 3 or more the following manifestations appear:
 - PA - no change
 - AP - pressure from handling
 - Caudad - reduces the pain
- Acromioclavicular joint (AP, PA, caudad) - no pain, no limitations

6.4 Accessory Movement Tests: Cervico-Thoracic Spine

- Left unilateral grade 3 or more to T2 is the only accessory movement that causes the patient some discomfort greater than the normal tenderness etc. Patient stiffens up a bit and reports a bit of discomfort over the area. Does not reproduce or refer pain to shoulder.

Appendix 6: Pre-Encounter Questionnaire

Pre-Encounter Questionnaire

Student Name/Code: _____

Date: _____

1. When thinking about conducting this assessment, I feel:

Very relaxed 1 2 3 4 5 **Very anxious**

2. When thinking about conducting this assessment, I feel:

Very unconfident 1 2 3 4 5 **Very confident**

Appendix 7: Post-Encounter Questionnaire

Post-Encounter Questionnaire

Student Name/Code: _____

Date: _____

1. Throughout the patient assessment I generally felt:

Very relaxed 1 2 3 4 5 **Very anxious**

2. Throughout the patient assessment I generally felt:

Very unconfident 1 2 3 4 5 **Very confident**

3. I found the patient to be:

Very Unconvincing 1 2 3 4 5 **Very convincing**

4. The subjective (history) findings that were presented by the patient were:

Very Unconvincing 1 2 3 4 5 **Very Convincing**

5. The objective (physical) findings that were presented by the patient were:

Very Unconvincing 1 2 3 4 5 **Very Convincing**

6. The psychosocial features of the patient's case were:

Very Unconvincing 1 2 3 4 5 **Very Convincing**

Comments: _____

Appendix 8: Demographic Questionnaire

Demographic Questionnaire

1. Student Name/Code: _____

2. Age: _____

3. Gender M F

4. The number of patients I have observed with shoulder pathology* over the past 18 months is/are:

Enter approximate number _____

5. The number of patients I have assessed with shoulder pathology* over the past 18 months is/are:

Enter approximate number _____

6. The number of patients I have treated with shoulder pathology* over the past 18 months is/are:

Enter approximate number _____

*The term 'shoulder pathology' referred to in this questionnaire can be direct or indirect. Direct shoulder pathology refers to situations where the pathology is the result of problems localised to the shoulder complex. In-direct pathology refers to problems which manifest in the shoulder but are referred from somatic tissues elsewhere (eg. cervical-thoracic spine).

Appendix 9: Instructions for Recall Session

Instructions for the Retrospective Think Aloud Recall Session, adapted from (Embrey, Guthrie, White, & Dietz, 1996; Ericsson & Simon, 1993)

As you watch the videotape of yourself with the client you have just assessed, we will be conducting a procedure called, ‘think aloud’. During this process, I’d like you to watch the videotape and verbalise what you are thinking, using these guidelines:

1. Verbalise what you are thinking at this point in time, not what you may have been thinking during the treatment session.
2. do not try to overanalyse. Just talk about whatever comes to your mind.
3. Speak as continuously as possible. Say something at least every 5 seconds, even if only, “I’m drawing a blank.”
4. You may stop the videotape at any time by pressing the pause button on the remote control device.
5. You may rewind the videtape at any time by pressing the rewind button on the remote control device.
6. Periodically, I may encourage you to share your thoughts.

Appendix 10: Instructions for Warm-Up Sessions

Instructions for the Retrospective Think Aloud Warm Up Sessions (Ericsson & Simon, 1993) p. 378.

In this experiment we are interested in what you think about when you find answers to some questions that I am going to ask you to answer. In order to do this I am going to ask you to **THINK ALOUD** as you work on the problem given. What I mean by think aloud is that I want you to tell me **EVERYTHING** you are thinking from the time you first see the question until you given an answer. I would like you to talk aloud **CONSTANTLY** from the time I present each problem until you have given your final answer to the question. I do not want you to try to plan out what you say or try to explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time I will ask you to talk. do you understand what I want you to do?

Good, now we will begin with some practice problems. First, I want you to multiply these two numbers in your head and tell me what you are thinking as you get an answer.

“What is the result of multiplying 24×36 ?

Good, now I want to see how much you can remember about what you were thinking from the time you read the question until you gave the answer. We are interested in what you actually can **REMEMBER** rather than what you think you must have thought. If possible I would like you to tell about your memories in the sequence in which they occurred while working on the question. Please tell me if you are uncertain about any of your memories. I do not want you to work on solving the problem again, just report all that you can remember thinking about when answering the question. Now tell me what you remember.

Good, now I will give you two more practice problems before we proceed with the main experiment. I want you to do the same thing for each of these problems. I want you to think aloud as before as you think about the question, and after you have

answered it I will ask you to report all that you can remember about your thinking.
Any questions?

Here is your next problem.

“How many windows are there in your parent’s house?”

Now tell me all that you can remember about your thinking.

Good, now here is another practice problem. Please think aloud as you try to answer it. There is no need to keep count., I will keep track for you.

“Name 20 animals”

Now tell me all that you can remember about your thinking.

Appendix 11: Data Screening

Data Screening (Coakes & Steed, 1997)

Testing for Normality

All data was examined for normality by exploring the following SPSS tools

1. Stem and leaf plots
2. Boxplots
3. Normal probability plot
4. Detrended normal plot
5. Kolmogorov-Smirnov with Lilliefors, Shapiro-Wilks (for samples < 50)

(if * $p < .05$ then normality is not assumed)
6. Skewness - value of 0 indicates an exact normal curve; + values suggest a positive skew; and - values a negative skew.
7. Kurtosis - value of 0 indicates an exact normal curve; + values suggest a positive skew; and - values a negative skew.

(items 1 - 4: not illustrated)

Assumptions Underlying t-Tests

The following assumptions were followed with respect to the administration of t-tests in this study

Generic Assumptions

1. The scale of measurement for all data was at the interval or ratio level of measurement.
2. Scores that were obtained were obtained by random sampling from the population of interest

3. Scores are normally distributed in the population

Independent samples: specific assumptions

1. Independence of groups: subjects appear in only one group and these groups are unrelated
2. Homogeneity of variance: the groups come from populations with equal variances (Levene test for equality of variances - **If this test is nonsignificant, $p > .05$, then there are no significant differences between the variances of the groups**)

Paired Samples: specific assumption

Normality of population difference scores should be normally distributed, providing the sample size is not too small (30+), violations of this assumption are of little concern.

Assumptions for ANOVA

1. Population Normality: populations from which the samples have been drawn should be normal. This can be evaluated using normality statistics such as skewness and Shapiro-Wilks ($n < 50$).
2. Homogeneity of Variance: scores in each group should have homogenous variances.

Assumptions for Testing the Correlation Coefficient

1. Normality - the scores within each variable should be normally distributed (see earlier data checks)
2. Homoscedasticity - the variability in scores for one variable is roughly the same at all values of the other variable. That is, it is concerned with how the scores cluster uniformly about the regression line. Scores, in this case, cluster uniformly around the regression line.

Missing Data and Outliers

There are no missing data in the data sets. Outliers have been included in the analysis.

COMFORT AND CONFIDENCE

df = 62	Kolmogorov-Smirnova	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test)
Pre-comfort (anxiety)	.000*	.444	-.326	.129
Post-comfort (anxiety)	.000*	.000	-.507	.607
Pre-confidence	.000*	.023	-.467	.729
Post-confidence	.000*	.074	-.018	.146

- The Kolmogorov-Smirnov measures do not suggest a normal distribution (if * $p < .05$ then normality is not assumed).
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if $p > .05$)
- The sample size is greater than 30 so violation of 'normality of population difference scores' is acceptable for paired sample t-testing.

HISTORY, PHYSICAL EXAMINATION AND ACIRS SCORES

df = 41	Shapiro-Wilks (for samples <50)	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
History Checklist	.017*	-1.107	2.699	.211
Physical Examination Checklist	.030*	-.315	-1.099	.104
ACIRS	.010*	-1.146	.623	.015*
Total Score	.039*	-.897	.805	.257

- The Shapiro-Wilk's measures do not suggest a normal distribution (if * p < .05 then normality is not assumed).
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05) except for the ACIRS.

TIME, THOROUGHNESS AND EFFICIENCY DATA

df = 41	Shapiro-Wilks (samples <50)	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
Time to complete History	.246	.123	-.981	.747
Time to complete Phys. Exam.	.083	.326	-.761	.889
Time to complete Tot. Encounter	.143	.362	-.749	.950
Thoroughness History	.017*	-1.107	2.699	.211
Thoroughness Phys. Exam	.030*	-.315	-1.099	.104
Thoroughness Tot. Encounter	.072	-.799	.417	.124
Efficiency History	.502	.360	-.310	.853
Efficiency Physical Examination	.216	.163	-.528	.092
Efficiency Total Encounter	.480	.366	-.308	.652

- The Shapiro-Wilk's measures suggest a normal distribution except for thoroughness of history and physical examination (if * p < .05 then normality is not assumed).
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05).

POST ENCOUNTER QUESTIONNAIRE DATA

df = 62	Kolmogorov-Smirnova	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
PEQ diagnosis	.000*	-1.108	.826	.074
PEQ management	.200	.358	.129	.508
PEQ history key features	.001*	-.542	.747	.323
PEQ physical exam key features	.000*	-.636	.142	.086
PEQ total	.200	-.383	.045	.019*

- The Kolmogorov-Smirnov measures do not suggest a normal distribution for diagnosis and the key features sections of the PEQ (if * p < .05 then normality is not assumed). The management and overall score categories indicate a normal distribution.
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05) except for the overall PEQ score.

LOW STUDENT DATA

df = 30	Shapiro-Wilks (samples <50)	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
PEQ diagnosis	.010*	-.961	.384	.007*
PEQ management	.956	.203	.229	.711
PEQ history key features	.017*	.066	-1.022	.940
PEQ physical exam key features	.015*	-.702	-.426	.041*
PEQ total	.363	-.451	.313	.056

- The Shapiro-Wilk's measures suggest a normal distribution except PEQ diagnosis and the key features section (if * p < .05 then normality is not assumed).
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05) except for diagnosis and physical examination key features.

HIGH STUDENT DATA

df = 32	Shapiro-Wilks (samples <50)	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
PEQ diagnosis	.010*	-1.112	1.741	.997
PEQ management	.076	.383	.015	.753
PEQ history key features	.048*	-.650	.822	.368
PEQ physical exam key features	.211	-.578	1.236	.446
PEQ total	.452	-.304	-.176	.647

- The Shapiro-Wilk's measures suggest a normal distribution except for PEQ diagnosis and the history key features section (if * p < .05 then normality is not assumed).
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05).

Measures Within the RPC Group

df = 18	Shapiro-Wilks (samples < 50)	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
Time History	.440	-.075	-1.087	.600
Time Physical Examination	.958	.121	-.525	.212
Time Total	.793	.106	-.838	.089
History Checklist	.335	-.263	-.410	.994
Physical Examination Checklist	.088	-.888	.470	.251
ACIRS	.010*	-1.755	5.141	.164
History and Physical Examination Checklist	.199	-1.027	1.456	.182
All Checklists	.063	-1.218	2.629	.115
Thoroughness	.199	-1.027	1.456	.182
Efficiency	.249	-.003	-1.302	.110
PEQ diagnosis	.010*	-1.170	2.123	.364
PEQ management	.070	.403	-.021	.194
PEQ history key features	.014*	-.780	.774	.194
PEQ physical examination key features	.010*	-1.038	1.459	.838
PEQ total score	.656	.131	.377	.993

- The Shapiro-Wilk's measures suggest a normal distribution except for the ACIRS, PEQ diagnosis and key features sections of the PEQ (if * p < .05 then normality is not assumed).
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05).

Measures Within the IND Group

df = 18	Shapiro-Wilks (samples < 50)	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
Time:Hx	.471	.386	-.562	.508
Time:Phys.Ex.	.060	.787	-.150	.619
Time: Total	.249	.854	.311	.682
HC Score	.091	-1.176	2.550	.688
PEC Score	.296	.297	-1.169	.360
ACIRS Score	.047*	-.689	-.733	.797
HC+PEC Score	.935	-.421	.062	.803
Total Score	.486	-.605	.483	.975
Mean Thorough.	.935	-.421	.062	.803
Mean Efficiency	.241	.220	-.668	.112
PEQ Diagnosis	.056	-.601	-.500	.107
PEQ Management	.467	-.091	-.636	.898
PEQ Key Features Hx	.016*	.634	-.291	.821
PEQ Key Features PEx.	.340	-.053	-.772	.280
PEQ Total	.173	-.177	-1.287	.993

- The Shapiro-Wilk's measures suggest a normal distribution except for the ACIRS and the PEQ Key Features History section (if * p < .05 then normality is not assumed).
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05).

Grapevine Effects

df = 37	Shapiro-Wilks	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
History Checklist	.017*	-1.107	2.699	.290
Physical Examination Checklist	.030*	-.315	-1.099	.621
History and Physical Exam. Checklists	.072	-.799	.417	.847

- The Shapiro-Wilk's measures suggest a normal distribution only for the History+Physical Examination Checklist Score (if * p < .05 then normality is not assumed).
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05).

df = 58	Kolmogorov-Smirnov	Skewness	Kurtosis	Homogeneity of Variance b/t Groups (Levene's Test) * p < .05
PEQ Total	.200	-.386	.034	.675

- The Kolmogorov-Smirnov measure suggest a normal distribution (if * p < .05 then normality is not assumed)
- Skewness and Kurtosis values and data screening of plots appear within normal limits.
- There is homogeneity of variance across the two groups (if p > .05)