

**Development of the Assessment of Physiotherapy Practice - A standardised
and validated approach to assessment of professional competence in
physiotherapy.**

Megan B Dalton BPhy, MAppSc

This thesis is submitted in fulfillment of the requirements of the Degree of Doctor of
Philosophy, Monash University

Department of Physiotherapy
School of Primary Health Care
Faculty of Medicine, Nursing and Health Sciences

February 2011

“In medio stat veritas”

Quintus Horatius Flaccus (December 8, 65 BC – November 27, 8 BC)

Table of Contents

List of Tables	x
List of Figures	xii
Declaration.....	xiv
Glossary of Terms	xv
Acknowledgements.....	xvii
Summary of Thesis Proposal.....	xix
Publication and presentations arising from research reported in this thesis	xxiv
Submitted Journal Articles.....	xxiv
Book Chapters.....	xxiv
Conference Presentations	xxiv
Reports.....	xxv
1. Chapter One: A systematic review of instruments for the assessment of professional competence of physiotherapy students.....	1
1.1 Introduction.....	1
1.2 Assessment of workplace-based clinical performance: evidence of validity	2
1.2.1 Validity evidence based on test content	4
1.2.2 Validity evidence based on internal structure.....	4
1.2.3 Validity based on relations to other variables.....	5
1.2.4 Validity based on response processes	5
1.2.5 Validity based on the consequences of testing (educational impact).....	6
1.2.6 Additional sources of validity evidence: acceptability and costs	6
1.2 Method.....	7
1.2.1 Identification and selection of studies.....	7
1.3 Data extraction.....	9
1.4 Results	14
1.4.1 Flow of papers through the review	14
1.4.2 Description of instruments	16
1.4.3 Validity evidence based on test content	16
1.4.4 Validity evidence based on response processes.....	18
1.4.5 Validity evidence based on internal structure.....	18
1.4.6 Validity evidence in relation to other variables.....	20
1.4.7 Validity evidence based on consequences of testing (educational impact)	21
1.4.8 Additional sources of validity evidence: acceptability and costs	22
1.5 Discussion.....	22

1.5.1	Implications for practice	27
1.5.2	Limitations of the review	28
2.	Chapter Two: The design and development of an instrument to assess professional competence in physiotherapy students	29
2.1	Introduction.....	29
2.1.1	Developing instruments to measure performance in the clinical context.....	31
2.2	Method.....	32
2.2.1	Construct Mapping (Step 1).....	33
2.2.2	Items design (Step 2)	34
2.2.3	Outcome space (scoring system) (Step 3)	35
2.2.4	Measurement model (Step 4).....	47
2.3	Action Research: Synthesis of instrument development.....	50
2.4	Instrument development: quality assurance processes	51
2.5	Instrument development: Summary	54
3.	Chapter Three: Development of the Assessment of Physiotherapy Practice (APP) instrument	55
3.1	Introduction.....	55
3.2	Methods (Part 1): Instrument development phase	55
3.2.1	Project Team and funding.....	55
3.2.2	Aims of the Research	55
3.2.3	Construct Mapping	56
3.2.4	Item design.....	56
3.2.5	Outcome space (scoring system).....	57
3.2.6	Measurement model	58
3.3	Methods (Part 2): Consultation phase	58
3.3.1	Focus groups	59
3.4	Results (Part 1) Instrument development phase	63
3.4.1	Construct Mapping	63
3.4.2	Item design.....	65
3.4.3	Outcomes space.....	65
3.5	Results (Part 2) Consultation phase (focus groups)	69
3.5.1	Content analysis.....	69
3.5.2	Revised instrument (version 2) for pilot trial.....	70
3.5.3	Mapping the standards.....	74
3.6	Discussion.....	74

4.	Chapter Four: Pilot Trial - Quantitative evaluation	77
4.1	Introduction.....	77
4.2.	Method	77
4.2.1	Participants – students and clinical educators	77
4.2.2	Pilot trial testing procedure – prior to commencement of clinical unit.....	78
4.2.3	Pilot trial testing procedure – during the clinical unit.....	79
4.2.4	Pilot trial testing procedure – Data management and analysis on completion of the clinical unit.....	80
4.2.5	Rasch analysis	80
4.3.	Results.....	84
4.3.1	Participants’ characteristics	84
4.3.2	Characteristics of item and instrument scoring.....	85
4.3.3	Rasch analysis: Model	87
4.3.4	Rasch analysis: Overall Model Fit	87
4.3.5	Overall Item and Person Fit	87
4.3.6	Individual Item and Person Fit	87
4.3.7	Threshold ordering of polytomous items	88
4.3.8	Targeting	89
4.3.9	Hierarchy of item difficulty	90
4.3.10	Person separation index (PSI)	90
4.3.11	Dimensionality	90
4.4	Discussion.....	91
4.5	Chapter Summary.....	94
5.	Chapter Five: APP Pilot Trial - Qualitative evaluation.....	95
5.1	Introduction.....	95
5.2	Method.....	95
5.2.1	Focus groups	96
5.2.2	Presentations	96
5.2.3	Data analysis	97
5.3	Results	98
5.3.1.	Participants	98
5.3.2	Analysis of the manifest content (quantitative assessment)	100
5.3.3	Analysis of the latent content (qualitative assessment)	103
5.3.4	Definition of entry level / passing standard performance	106
5.3.5	Addition of global rating scale (GRS)	107

5.3.6	APP (version 3) for use in Field Test One.....	108
5.4	Discussion.....	114
6.	Chapter Six: Field Test One - Qualitative evaluation	117
6.1	Introduction.....	117
6.2	Methods : preparation for Field Test One	117
6.2.1	Development of training resources	117
6.2.2	Development of demographic data forms and feedback questionnaires.....	119
6.2.3	Recruitment of participants.....	120
6.3	Methods Stage Two: during Field Test One	123
6.3.1	Field Test One procedure – during the clinical education unit	123
6.3.2	Teleconferences with clinical educators.....	123
6.3.3	Think aloud interviews.....	126
6.4	Methods Stage 3: on completion of Field Test One.....	129
6.4.1	Data management and analysis.....	129
6.4.2	Focus groups conducted following Field Test One	129
6.5	Results: Qualitative evaluation Field Test One.	131
6.5.1	Development of training resources	131
6.5.2	Demographic data form and feedback questionnaire.....	131
6.5.3	Participant characteristics	131
6.5.4	Participant training: Workshops, and teleconferences	133
6.5.5	Think aloud interviews.....	133
6.5.6	Feedback questionnaires	137
6.5.7	Demographic form: Clinical educator experience	140
6.5.8	Focus groups conducted following Field Test One	140
6.6	Discussion.....	145
6.7	Chapter Summary.....	148
7.	Chapter Seven: Field Test One - Quantitative evaluation	149
7.1	Introduction.....	149
7.2.	Method	149
7.2.1	Participants – students and clinical educators	149
7.2.2	Field testing procedure	150
7.2.3	Data management and analysis.....	150
7.3.	Results.....	152
7.3.1	Participant characteristics	152
7.3.2	Characteristics of item and instrument scoring.....	152

7.3.3	Factor analysis.....	154
7.3.4	Rasch analysis: Overall Model Fit	157
7.3.5	Overall Item and Person Fit	157
7.3.6	Individual Item and Person Fit	158
7.3.7	Threshold ordering of polytomous items	161
7.3.8	Targeting	162
7.3.9	Hierarchy of item difficulty	164
7.3.10	Person separation index	166
7.3.11	Differential item functioning (DIF).....	166
7.3.12	Dimensionality	168
7.4	Discussion.....	169
7.5	Actions arising following Field Test One	173
7.6	Chapter Summary.....	175
8.	Chapter Eight: Field Test Two - Qualitative evaluation.	176
8.1	Introduction.....	176
8.2	Methods	176
8.2.1	Stage 1: Preparation for Field Test Two.....	176
8.2.2	Stage 2: during Field Test Two	179
8.2.3	Stage 3: on completion of Field Test Two.....	180
8.5	Results: Qualitative evaluation Field Test Two.	182
8.5.1	Participant characteristics	182
8.5.2	Participant training: Workshops, and teleconferences	184
8.5.3	Feedback questionnaires	185
8.5.4	Focus groups conducted following Field Test Two.....	187
8.6	Discussion.....	194
8.7	Chapter Summary.....	196
9.	Chapter Nine: Field Test Two - Quantitative evaluation	197
9.1	Introduction.....	197
9.2	Validity evidence based on relations to other variables.....	197
9.2.1	Convergent and discriminant evidence	197
9.2.2	Developmental progression in competency.....	198
9.3	Method.....	198
9.3.1	Field Test Two	198
9.3.2	Validity evidence based on relations to other variables	198
9.3.3	Data management and analysis.....	200

9.4	Results	201
9.4.1	Participants' characteristics – Field Test Two.....	201
9.4.2	Characteristics of item and instrument scoring.....	201
9.4.3	Characteristics of orthopaedic examination results.....	203
9.4.4	Characteristics of APP scores across six clinical units (n=57)	205
9.4.5	Factor analysis.....	206
9.4.6	Rasch analysis: Model	209
9.4.7	Rasch analysis: Overall Model Fit	209
9.4.8	Overall Item and Person Fit	210
9.4.9	Individual Item and Person Fit	210
9.4.10	Threshold ordering of polytomous items	213
9.4.11	Targeting	214
9.4.12	Hierarchy of item difficulty	215
9.4.13	Person separation index	217
9.4.14	Differential item functioning (DIF).....	217
8.3.15	Dimensionality	218
8.3.16	Relationship of global ratings to person measures	219
9.4	Discussion.....	220
9.5	Actions arising following Field Test Two	225
9.6	Chapter Summary.....	225
10.	Chapter Ten: Reliability.....	229
10.1.	Introduction.....	229
10.1.1	Establishing Reliability	230
10.2	Method	236
10.2.1	Study design.....	236
10.2.2	Recruitment of participants.....	236
10.2.3	Ethics approval.....	237
10.2.4	Trial preparation	237
10.2.5	Trial procedure – during the clinical unit.....	238
10.2.6	Trial procedure – on completion of the clinical unit	238
10.2.7	Data management and analysis.....	239
10.3	Results.....	239
10.3.1	Participant characteristics	239
10.3.2	Relationship between raters.....	240
10.3.3	Paired t-test	242

10.3.4	Intraclass correlation coefficient ICC(2,1).....	242
10.3.5	Standard Error of Measurement (SEM).....	244
10.3.6	Minimal Detectable Change (MDC).....	244
10.3.7	Bland Altman analyses.....	244
10.4	Discussion.....	247
11.	Chapter Eleven: Validity, future research directions and summary.....	253
11.1.	Introduction.....	253
11.2	Sources of validity evidence: framework for instrument development.....	254
11.3	Validity evidence based on content.....	258
11.4	Validity evidence based on internal structure.....	259
11.5	Validity based on relations to other variables.....	260
11.5.1	Convergent and discriminant evidence.....	260
11.5.2	Predictive evidence.....	261
11.5.3	Developmental Progression in Competency.....	261
11.6	Validity based on response processes.....	262
11.6.1	Think aloud interviews and focus groups.....	262
11.6.2	Rating scale analysis.....	263
11.6.3	Rater and student training.....	263
11.6.4	Item hierarchy.....	264
11.7	Validity based on the consequences of testing (educational impact).....	264
11.7.1	Impact on student learning: benefits to learning and any unintended negative consequences.....	265
11.7.2	Method of determining passing score determining pass/fail score and estimation of the standard error of measurement.....	266
11.8	Additional sources of validity evidence.....	267
11.9	Future research directions.....	268
11.10	Chapter Summary.....	274
12.	References.....	275
13.	Appendices (refer to volume 2).....	292

List of Tables

Table 1.1: Characteristics of the instruments identified in the review	11
Table 1.2: Summary of evidence for validity, reliability, acceptability and costs	13
Table 2.1: Summary work-based assessment instrument development: quality assurance processes.....	53
Table 3.1: Decisions to be made in relation to construction of a rating scale	66
Table 3.2: Coding guide for content analysis of focus groups pre pilot trial	69
Table 3.3: Summary of results of focus groups 1 and 2.	71
Table 3.4: Summary changes to CAPS (version 1) prior to pilot trial	72
Table 4.1: Descriptive statistics (n=295 completed assessments)	86
Table 4.2: Item order, average location and standard error (SE) from least to most difficult of the 20 items	90
Table 5.1: Focus group participants (after pilot trial and prior to Field Test One).....	99
Table 5.2: Presentations on APP results (after pilot trial before field test 1)	100
Table 5.3: Initial concept ranking from focus groups	102
Table 5.4: Initial concept ranking of co-occurrences of the word 'student'	104
Table 5.5: Summary of focus group data following pilot trial	109
Table 5.6: Modifications to APP (version 2) following pilot trial.....	112
Table 6.1: Field Test One: Participating universities.....	122
Table 6.2: Sample think aloud interview data collection form.....	128
Table 6.3: Field Test One participant and placement characteristics	132
Table 6.4: Field Test One clinical educator training.....	133
Table 6.5: Coding guide for content analysis of think aloud interviews	134
Table 6.6: Clinical educator feedback on APP	137
Table 6.7: Student feedback on APP.....	138
Table 6.8: Summary student focus group results Field Test One	144
Table 7.1: Descriptive statistics Field Test One (n=729)	153
Table 7.2: Component Matrix Field Test One	155
Table 7.3: Factor analysis parallel analysis Field Test One	156
Table 7.4: Component matrix	157
Table 7.5: Individual item fit of 20 APP items to the Rasch model: Sample 1 (n=390) and sample 2 (n=340)	160
Table 7.6: Uniform and non-uniform DIF statistics for all APP items for student gender for sample 1 (n=390)	166
Table 7.7: Modifications to APP (version 3) following Field Test One.....	174
Table 8.1: Field Test Two: participant characteristics (n=644).....	183
Table 8.2: Field Test Two clinical educator training	184
Table 8.3: Clinical educator feedback on APP (n=222)	185
Table 8.4: Preferred training options.....	186
Table 8.5: Student feedback on APP (n=251)	186
Table 8.6: Student feedback questionnaire results – dichotomous questions	187
Table 8.7: Coding guide for content analysis of focus groups	188
Table 8.8: Summary clinical educator focus group results Field Test Two.....	191
Table 8.9: Summary student focus group results Field Test Two.....	193
Table 9.1: Descriptive statistics Field Test Two (n=644).....	202
Table 9.2: Descriptive statistics of orthopaedic examination results (n=94)	203

Table 9.3: Correlations between different orthopaedic examination formats	204
Table 9.4: Descriptive statistics of six clinical blocks (n=57).....	205
Table 9.5: Component Matrix Field Test Two.....	207
Table 9.6: Factor analysis parallel analysis Field Test Two	208
Table 9.7: Component matrix	209
Table 9.8: Individual item fit of 20 APP items to the Rasch model: Sample 1 (N=326) and sample 2 (n=318)	212
Table 9.9: Uniform and non-uniform DIF statistics for all APP items for student gender	218
Table 9.10: Modifications to APP (version 4) following Field Test Two	226
Table 9.11: Summary qualitative and quantitative data following pilot and field testing....	227
Table 10.1: Demographics for participants in the inter-rater reliability trial.....	240
Table 10.2: Percent agreement between raters on item and global rating scores.....	241
Table 10.3: ICC 2,1 for items, domains of practice, GRS and total score on APP.....	243
Table 10.4: Results of previous inter rater reliability trials conducted in clinical environment relocate to after discussion commences	246
Table 11.1: Sources of validity evidence.....	255

List of Figures

Figure 1.1: Comparison of validity frameworks (adapted from Baartman et al 2007).	3
Figure 1.2: Flow of papers through the review	15
Figure 2.1: The instrument development cycle through the four building blocks (adapted from Wilson, 2005 p. 19)	33
Figure 2.2: Construct map for physical functioning (adapted from Wilson, 2005 p. 31)	34
Figure 2.3: Examples of rating scales: VAS, DVAS, Adjectival and Likert.....	39
Figure 2.4: Examples of rating scales: BAS, BOS and GRS.....	41
Figure 2.5: Iterative research framework	52
Figure 3.1: Construct map.....	64
Figure 4.1: Threshold map for items 1 - 20.....	88
Figure 4.2: Category probability curves showing ordered thresholds for item 11 (Identifies and prioritises patient's/client's problems). Locn=location; FitRes= Fit Residual; ChiSq[Pr]=chi-square probability	88
Figure 4.3: Person item distribution graph for total APP scale	89
Figure 5.1: Concept map development in Leximancer (Kivunja 2009).....	98
Figure 5.2: Concept map displaying frequency of concepts arising from focus group transcripts	101
Figure 5.3: Concept map showing co-occurrences of the word 'student'	103
Figure 6.1: Flow chart of Field Test One procedure	125
Figure 7.1: Scree plot	155
Figure 7.2: Threshold map of APP 20 items in sample 1 n=390	161
Figure 7.3: Category probability curves for item 15 (Is an effective educator) in data from sample 2 (n=340). Locn=location; FitRes= Fit Residual; ChiSq[Pr]=chi-square probability... ..	162
Figure 7.4: Person-item threshold distribution graph for sample 1 (n=390).	163
Figure 7.5: Person-item threshold distribution graph for sample 2 (n=340).	163
Figure 7.6: Logit location of APP items in two samples (sample 1 n=390; sample 2 n=340) ..	164
Figure 7.7: Plot of person logit location and raw APP score (sample 1)	165
Figure 7.8: Plot of person logit location and raw APP score (sample 2)	165
Figure 7.9: Differential item functioning graph of female and male students for item 6 (written communication) sample 1 (n=390).....	167
Figure 7.10: Differential item functioning graph of female and male students for item 5 (communication - verbal/non-verbal) sample 1 (n=390)	167
Figure 9.1: Scatter plot of total scores for orthopaedic practical exam and clinical unit	205
Figure 9.2: Change in mean APP scores across six clinical blocks	206
Figure 9.3: Scree plot Field Test Two.....	207
Figure 9.4: Threshold map of APP 20 items in sample 1 (n=326).....	213
Figure 9.5: Category probability curves for item 4 in sample 2 (n=318)	213
Figure 9.6: Person-item threshold distribution graph for sample 1 (n=326)	215
Figure 9.7: Person-item threshold distribution graph for sample 2 (n=318)	215
Figure 9.8: Logit location of APP items in two samples for Field Test Two.....	216
Figure 9.9: Plot of person logit location and raw APP score (sample 1)	216
Figure 9.10: Plot of person logit location and raw APP score (sample 2)	217
Figure 9.11. Scatter plot of Global rating scores (overall rating of competence) against Rasch person location logit score for sample 1 (n=326) in Field Test Two.....	220
Figure 10.1: Scatterplot of APP scores for rater 1 and rater 2	242

Figure 10.2: Plot of the differences between raters' marks against the means of raters' marks for the total score out of 80 (n=60 assessments). The x-axis bisects the y-axis at the mean difference between raters and the upper and lower lines represent the 95% limits of agreement.....245

Declaration

The work contained in this thesis is, to the best of my knowledge and belief, original, except as acknowledged in the text. The material has not been submitted, either in whole or in part, for a degree at this or any other university. All the raw data pertaining to the studies reported in this thesis, as well as the analyses, have been retained and are available on request. All research procedures were approved by the Monash University Standing Committee for Ethics in Research in Humans.

This thesis has been prepared based on the style recommended in the Publication Manual of the American Psychological Association (5th edition). Spelling used in this thesis conforms to Australian English.

.....

Megan Dalton

.....

Date

Copyright

Notice 1

Under the Copyright Act 1968, this thesis must be used only under the normal conditions of scholarly fair dealing. In particular no results or conclusions should be extracted from it, nor should it be copied or closely paraphrased in whole or in part without the written consent of the author. Proper written acknowledgement should be made for any assistance obtained from this thesis.

Notice 2

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

Glossary of Terms

ACOPRA	Australian Council of Physiotherapy Regulating Authorities
AMEE	Association for Medical Education in Europe
AMED	Allied and Complementary Medicine
ANZ	Australian New Zealand
APC	Australian Physiotherapy Council
APP	Assessment of Physiotherapy Practice
BAS	Behaviourally Anchored Scales
BEI	British Education Index
BEME	Best Evidence Medical Education
BMACS	(Blue) Mastery and Assessment of Clinical Skills
BOS	Behavioral Observation Scale
CAF	Common Assessment Form
CAPS	Clinical Assessment of Physiotherapy Skills
CIET	Clinical Internship Evaluation Scale
CINAHL	Cumulative Index on Nursing and Allied Health Literature
COMPASS TM	National speech pathology competency based assessment tool
CPI	Clinical Performance Instrument
CTT	Clinical Test Theory
DIF	Differential Item Function
DVAS	Discrete Visual Analog Scale
ECC	Evaluation of Clinical Competency
ERIC	Education Resource Information Centre
GPA	Grade Point Average
GRS	Global Rating Scale

HaPI	Health and Psychosocial Instruments
ICC	Intraclass Coefficient
IRT	Item Response Theory
MDC ₉₀	Minimal Detectable Change at 90% confidence interval
NPTE	National Physical Therapy Examination
OSCE	Observed Structured Clinical Examination
PCA	Principle Components Analysis
PEDRO	Physiotherapy Evidence Database
PIs	performance Indicators
PTMACS	Physical Therapy Manual for the Assessment of Clinical Skills
PRISMA	Preferred Reporting Items for Sytematic Reviews and Meta-Analyses
PSI	Person Separation Index
RMM	Rasch Measurement Model
RUMM2020	Software for Rasch analysis
SCCS	Standard Clinical Competency Scale
SEM	Standard Error of Measurement
SPSS	Statistical Package for the Social Sciences
USA	United States of America
VAS	Visual Analogue Scale
WCPT	World Confederation of Physical Therapists

Acknowledgements

Completing my PhD has been taken several years and many people have supported me during this journey. While words cannot convey the depth of my gratitude, there are many people I wish to acknowledge.

To my supervisor Professor Jenny Keating, I offer my sincere thanks for her unflagging enthusiasm, intellectual curiosity and constructive feedback. Even when I felt completely overwhelmed by the task, she did not falter but rather rose to the challenge and urged me on. She taught me the value of patience, tenacity and belief in myself. I would also like to acknowledge the support provided by my associate supervisor, Associate Professor Megan Davidson from La Trobe University. Megan provided the yin to Jenny's yang and for that I am forever grateful. Dr Natalie de Morton and Dr Julie Pallant assisted me in my journey to understanding Rasch analysis.

I would like to thank the university academic and clinical staff, physiotherapy clinical educators and physiotherapy students across Australia and New Zealand for their enthusiasm, support and hard work throughout the period of this research. The Clinical Education Managers from each of the Physiotherapy university programs throughout Australia and New Zealand who formed the reference group, and the Council of Deans of Physiotherapy, Australia and New Zealand (CPDANZ) also contributed valuable expertise and support. I wish to thank Wendy Harris for her work as research assistant and administrative manager. The library staff at Griffith University were always generous with their time and support.

Dr Sue McAllister from the University of Sydney/Flinders University, who has conducted similar project with Speech Pathology Australia, was a valuable resource and provided constructive suggestions and moral support. Thanks to Dr Liz Molloy, Wendy Nixon and physiotherapy students from Monash University who assisted in the production of the APP training DVD.

A special vote of thanks must go to the physiotherapy staff within the School of Physiotherapy and Exercise Science at Griffith University for providing me with the time to complete my PhD.

To Kieran, Mum and Dad, Bernardine, Ros F, Lawrence P, Luke and Clare thank you for your words of wisdom, encouragement, cups of coffee, regular reality checks and humour. To David Butler and Gwen Jull thank you for starting me on this journey many years ago. To Sally, your friendship, love and support revived me many times.

Above all the research team is indebted to the Australian Learning and Teaching Council for enabling this work through funding and support. Without financial aid, this research would still be a hope not a reality.

Summary of Thesis Proposal

Valid, reliable and standardised assessment formats and procedures, suited to application in the workplace, are important for meaningful and consistent assessment of the clinical performance of physiotherapy students. The choice of clinical assessment instruments for physiotherapy programs in Australia has typically been influenced by historical precedents and the personal experience of assessors rather than by the known strengths and weaknesses of an assessment instrument, a situation common to that observed in medical programs (E. D. Newble, Jolly, & Wakeford, 1994).

The Queensland Health Clinical Education Project (2005) acknowledged the variability of procedures and instruments for assessment of physiotherapy practices across different universities in Australia. At that time there were 16 entry-level physiotherapy programs in Australia, all accredited by the Australian Physiotherapy Council (APC). Each physiotherapy program was required to demonstrate that graduates met the performance standards outlined in the Australian Standards for Physiotherapy (2006). Despite each program having curriculum designed to meet the same standards, when this thesis commenced each physiotherapy program used unique clinical assessment forms and assessment criteria. The Queensland Health Clinical Education Project emphasised the diversity of assessment forms and supporting documentation as a substantial and unnecessary burden on assessors who were required to use multiple assessment instruments. In addition, the measurement properties of these assessment instruments were unknown, impacting on confidence in the reliability and validity of decisions based on these assessment approaches. As new physiotherapy programs commenced, this burden multiplied.

This thesis describes the development of a standardized assessment instrument to meet the needs of physiotherapy students and educators and provide valid and reliable measurements of clinical performance. The need for this research was identified by university-based physiotherapy programs across Australia and New Zealand, physiotherapy educators and supervisors, and the APC responsible for accreditation of physiotherapy programs within Australian universities. Funding was provided by the Australian Learning

and Teaching Council (formerly The Carrick Institute) to commence work on the development of an assessment instrument.

The research in this thesis is reported in chronological order with each phase informing subsequent steps. Streiner and Norman (2003) proposed that the first step in the development of a new instrument was to be fully informed of existing scales and the quality of such instruments prior to embarking on the development of a new instrument. This work began with a systematic review of methods used to assess professional competence in physiotherapy practice (Chapter One).

The systematic review found a number of reports of research into assessment of competence in physiotherapy practice; these varied in design and method quality (see Chapter One). Eight instruments developed to assess professional competence of physiotherapy students within the clinical environment were located. The review failed to identify convincing evidence sufficient to support the merits of one instrument above others.

In addition, investigation of the psychometric properties of these instruments was not performed utilising Item Response Theory (IRT) or the Rasch Measurement Model (RMM), rather employing the Classical Test Theory (CTT) approach. The thesis argues for the need to investigate instrument properties using IRT or RMM; these approaches offer substantial clinical and scientific advantage over traditional psychometric methods in the development and evaluation of rating scales, and in the analysis of rating scale data (Andrich, 1988; J. Hobart & Cano, 2009; Wilson, 2005; Wright, 1996a; Wright & Mok, 2000).

Chapter Two describes and defends the plan for instrument development. The first phase (Chapter Three) involved development of the assessment instrument content, format and processes. The research was guided by the Standards for Educational and Psychological Testing, (American Educational Research Association, 1999). The process of test design was based on the 'four building blocks' approach outlined by Wilson (2005) which comprised construct mapping, items design, outcome space and measurement model.

Once development of the first version of the instrument with the working title Clinical Assessment of Physiotherapy Skills (CAPS) was complete, cycles of action and reflection on outcomes (an action research approach) were utilised. The iterative research cycles included preliminary information gathering, instrument development, pilot trial / field test stages, and continuous refinement of the instrument based on evaluation throughout the different phases following recommendations for best practice in research of this nature (Coghlan & Brannick, 2001).

A pilot trial (Chapters 4 and 5) was conducted, using the instrument to assess 295 third and fourth year physiotherapy students. Rasch analysis of outcome data showed an overall fit to the Rasch model. The difficulty of the items was well matched to the abilities of the persons being assessed and the 5-level rating scale performed as intended. The results of the pilot trial supported the continuation of the research into field tests one and two.

The results of both field tests (Chapters 7 and 8) supported the findings of the pilot trial demonstrating that the APP data had adequate fit to the chosen measurement model (Rasch Partial Credit Model), the Person Separation Index demonstrated the scale was internally consistent discriminating between four groups of students with different levels of professional competence (0.96), the items were targeting the intended construct (professional competence) and the instrument demonstrated unidimensionality. Additionally differential item function (DIF) studies demonstrated there was no item bias in either field test for the variables: student age, gender and level of clinical experience, clinical educator age, gender and experience as an educator, facility type, and clinical area.

Qualitative data (Chapters 6 and 9) provided evidence of the acceptability of the instrument for use within the work place by educators and students. Further research investigating how educators were interpreting and scoring written communication and the impact on student learning of the assessment process was recommended. Ongoing evaluation and refinement of training methods and resources was also advocated.

The results of field testing provided data supported the validity of the APP instrument scores and acceptability of the instrument for use within the workplace. These data enabled

the final phase of research, investigation of inter-rater reliability, to proceed (Chapter Ten). Thirty pairs of clinical educators (60 independent educators) and 30 third and fourth year physiotherapy students from five universities participated in the reliability trial. Both correlational coefficients and metricated errors were estimated to provide a comprehensive analysis of the likely utility of APP scores and to enable score and change score interpretation. The Intraclass Correlation Coefficient 2,1 (two-way random effects model) for total APPs scores for the two raters was 0.92 (95% CI 0.84 to 0.96) and the ICC 2,1 for the global rating scale scores was 0.72 (95% CI 0.50 – 0.86). The 95% confidence band around a single score for this data was 6.5 APP points. With a scale width of 0 – 80, an error margin of 6.5 (95%CI) was considered acceptable. This error enables a high level of accuracy in ranking student performance as evidenced by test/retest correlation of .92.

For the APP the magnitude of change in scores required to conclude that real change has occurred is in the order of 7.8 points which compared favourably to other instruments used to assessment professional competence of physiotherapy students (Coote, et al., 2007; Meldrum, et al., 2008; Task Force for the Development of Student Clinical Performance Instruments, 2002).

Overall the physiotherapy clinical educators demonstrated a high level of reliability in the assessment and marking of physiotherapy students' performance on clinical placements when using the APP. This was found despite the variability anticipated due to different areas of practice, types of facilities and a spectrum of educator experience.

The final step in the research was to evaluate the evidence for validity of APP scores. Using the five sources of validity evidence presented in the American Educational Research Association (1999) standards, data from multiple sources was accumulated to establish the likely validity of interpretations made based on the instrument scores. The validity of scores for workplace-based professional competence awarded by educators to pre-entry level physiotherapy students using the APP was evaluated through Rasch analysis, parametric statistical evaluation, and qualitative data obtained from multiple sources. This approach enabled triangulation and reinforcement of decisions based on quantitative and qualitative

data obtained from multiple sources. The APP was found to have strong validity characteristics across all five sources of validity evidence as described in Chapter Eleven.

The APP was developed and applied within the constraints of a dynamic and unpredictable clinical environment. This is a key strength of the assessment instrument. The research has delivered an important benefit for physiotherapy education in that a single instrument with known validity and reliability is now available to replace the twenty-five distinct assessment forms formerly being used. To date, 17 out of 18 Universities in Australia and New Zealand have adopted the APP as the sole assessment form, and a further three new programs commencing within the next two years are also adopting the instrument.

This thesis has generated new evidence to support the development of the APP as a valid instrument for the workplace based assessment of professional competence of physiotherapy students, leading to standardised assessment at a national level.

Publication and presentations arising from research reported in this thesis

Submitted Journal Articles

Dalton M, Davidson M, Keating J.L. (2010). A systematic review of instruments for the assessment of professional competence of physiotherapy students. Currently under review: *Advances in Health Sciences Education*.

Book Chapters

Keating J, **Dalton M**, Davidson M (2009)

Title: Assessment in clinical education. In: *Clinical education in the health professions: an educator's guide*. Delany C, Molloy L (eds). Elsevier, Sydney.

Conference Presentations

1. **Dalton, M.**, Jolly B, Molloy E. (2010). Clinical assessment in the health professions: facilitating learning through the assessment process 14th Ottawa International Conference on Clinical Competence, 17-20 May, Miami, Florida, USA.
2. **Dalton, M.**, Keating, J. & Davidson, M., de Morton N. (2009) Development of a method to assess performance of entry level physiotherapists: the Assessment of Physiotherapy Practice (APP). ANZAME09 Bridging Professional Islands Conference, June 2009, Launceston, Australia
3. **Dalton, M.**, Keating, J. & Davidson, M. (2008). Development of the Assessment of Physiotherapy Practice (APP) Instrument: investigation of the psychometric properties using Rasch analysis. International Conference on Outcomes Measurement 2008, 11-13 September, Bethesda, Washington DC, USA,
4. **Dalton, M.**, Keating, J. & Davidson, M. (2008). Development of the Assessment of Physiotherapy Practice (APP) Instrument: results of Field Test One. APA National Congress 2008, 23-24 May, Perth, Australia
5. **Dalton, M.**, Keating, J. & Davidson, M. (2008). Development of the Assessment of Physiotherapy Practice (APP) Instrument: A standardised and valid approach to

assessment of clinical competence in physiotherapy. 13th Ottawa International Conference on Clinical Competence (Ozzawa), 5-8 March, Melbourne, Australia.

6. **Dalton, M.**, Keating, J. & Davidson, M. (2008). Development of the Assessment of Physiotherapy Practice (APP) Instrument: pilot trial results. 13th Ottawa International Conference on Clinical Competence (Ozzawa), 5-8 March, Melbourne, Australia.
7. **Dalton M** 2007. Developing the APP (Assessment of Physiotherapy Practice) – A standardized and validated approach to assessment of clinical competence in physiotherapy. Colloquium: Benchmarking student learning in the workplace – University of Newcastle centre for teaching and learning. June 2007

Reports

1. **Dalton M.**, Keating J., Davidson M. (2009, March). Development of the Assessment of Physiotherapy Practice (APP): A standardised and valid approach to assessment of clinical competence in physiotherapy. [Australian Learning and Teaching Council (ALTC) Final report PP6-28]. Brisbane: Griffith University. Available online at: <http://www.altc.edu.au/project-development-clinical-assessment-2006>
2. **Dalton M.**, Keating J., Davidson M. (2009, March). Development of the Assessment of Physiotherapy Practice (APP): A standardised and valid approach to assessment of clinical competence in physiotherapy. Educators resource Manual [Australian Learning and Teaching Council (ALTC) Final report PP6-28]. Brisbane: Griffith University. Available online at: www.altc.edu.au

1. Chapter One: A systematic review of instruments for the assessment of professional competence of physiotherapy students.

1.1 Introduction

Workplace-based learning is an essential component of all professional education programs. Typically assessment occurs across all levels of clinical competence and includes direct assessment of clinical practice. It is assumed that observed practice in the operational context is an indicator of likely professional performance (Wiggins 1989). Professional practice necessitates dealing with highly variable circumstances and assessments can be difficult to standardise across student groups (Rethans et al 2002). Controlled assessments such as Objective Structured Clinical Examinations (OSCEs) and the use of standardized patients have been developed in response to concerns regarding standardized and reliable measurement of student competencies. While assessment reliability may be enhanced by standardized testing, the validity of controlled examination procedures has been challenged because competence under controlled conditions may not be an adequate surrogate for performance under the complex and uncertain conditions encountered in usual practice (Southgate, et al., 2001).

In 1990, psychologist George Miller proposed a pyramidal hierarchy to describe assessment of clinical competence. From lowest to highest, the levels were defined as *knows*, *knows how*, *shows how* and *does*. For health professionals, ideal assessment procedures for the highest level of 'does' should facilitate evaluation of the complex domains of clinical performance in the context of the circumstances within which competence is desirable, that is, in the practice environment (Morton, Cumming, & Cameron, 2007; Norcini, Blank, Duffy, & Fortna, 2003). This level represents professional competence, defined by Epstein and Hundert (2002) as:

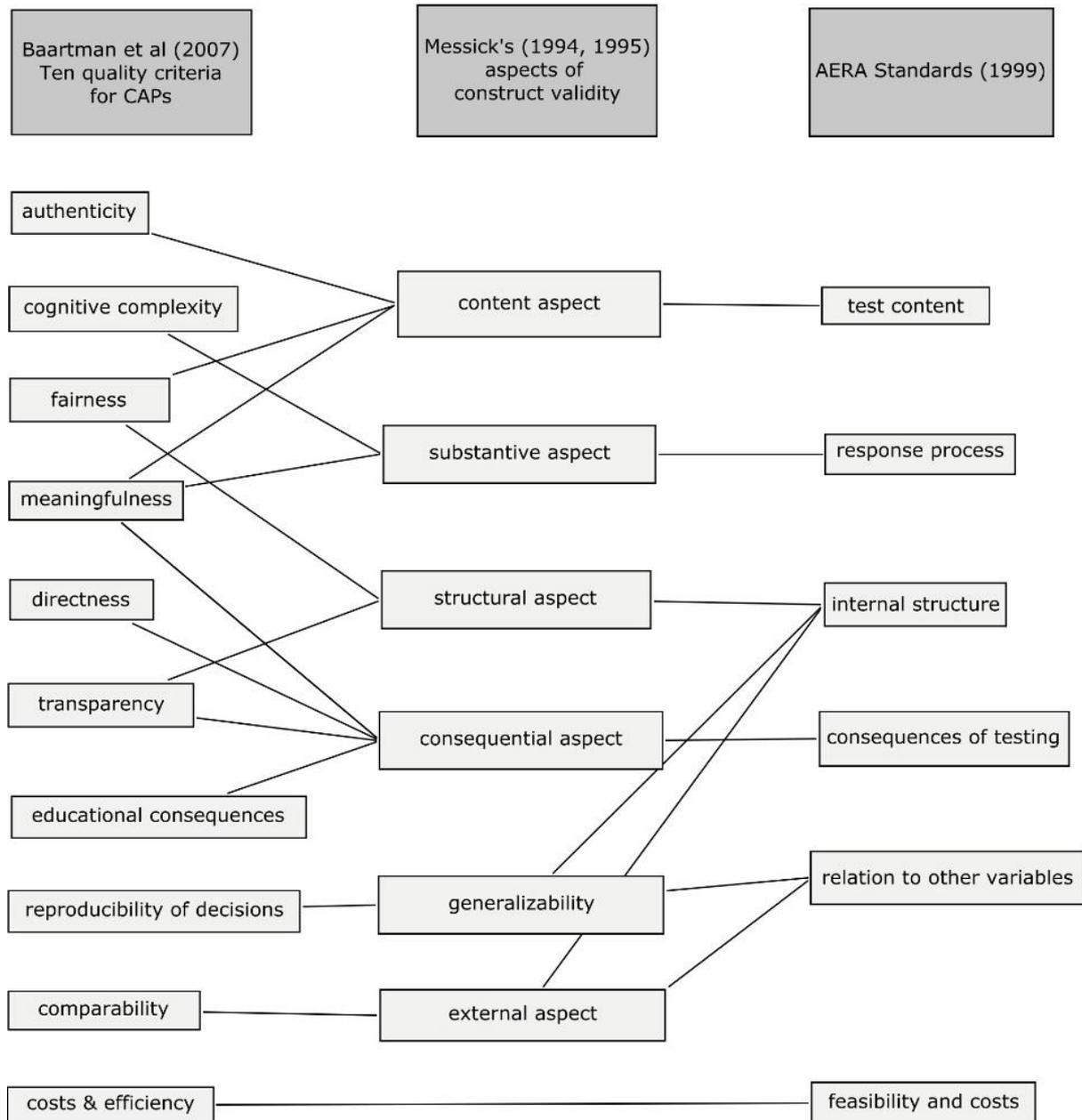
“The habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values and reflection in daily practice for the benefit of the individual and community being served.” (p 226)

Given the high stakes of summative assessments of clinical performance, assessment procedures should not only be feasible and practical within the clinical environment, but also demonstrate sufficient reliability and validity for the purpose (Epstein & Hundert 2002, Roberts et al 2006, Baartman 2007). Wass et al (2001a) considered the assessment of clinical competence to be a critically important international challenge, and one that must be based on sound measurement principles while ensuring that the learning process is supported and not threatened by the assessment approach.

1.2 Assessment of workplace-based clinical performance: evidence of validity

Currently, there are many approaches to establishing validity of assessment methods (American Educational Research Association, 1999; Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006, 2007; Fitzpatrick, Davey, Buxton, & Jones, 1998; Messick, 1989, 1994, 1995b; Streiner & Norman, 2003; Wilson, 2005; Wolfe & Smith, 2007a). Traditionally, validity was viewed as having three distinct primary facets: content, criterion and construct (Cronbach, 1955; Guion, 1980). Messick (1989) (1996) challenged this approach, proposing a unitary concept of validity with six interdependent aspects: content validity, substantial validity, structural validity, consequential validity, external validity and generalisability.

Baartman et al (2007) extended Messick's (1989) validity concept by developing a framework of ten quality criteria specifically designed to assess the complex combinations of knowledge, skills and attitudes required in competence based education. Figure 1.1 illustrates a qualitative comparison of Messick's construct validity framework with the quality assessment framework of Baartman et al and the approach recommended by the American Educational Research Association (1999). This figure highlights the overlap with Messick's validity categories.



(Legend: CAPs: competency assessment programs; AERA: American Education Research Association)

Figure 1.1: Comparison of validity frameworks (adapted from Baartman et al 2007).

The current Standards for Educational and Psychological Testing (1999) also embrace the integrated concept of validity proposed by Messick (1989) and recommend that evidence of validity be assembled from multiple sources to inform interpretations of instrument scores.

The Standards recommend that validity evidence be sought in five areas based on test content, response processes, internal structure, relations to other variables and consequences of testing.

1.2.1 Validity evidence based on test content

Important validity evidence can be:

“Obtained from an analysis of an instrument’s content and the relationship between the content and the construct it is intended to measure” (American Educational Research Association, 1999) (p11).

A transparent consensus approach to instrument development, with input from an appropriate spectrum of key stakeholders (e.g. national accrediting bodies, graduate employers, educators who use the assessment procedures, academics, students and a broad spectrum of practitioners) would provide evidence for validity based on test content.

Authenticity in assessment processes is also an integral aspect of validity evidence based on test content. Authentic assessment should assess the knowledge, attitudes and skills needed in the future workplace and promote quality learning in the practice context (Baartman, et al., 2007; Boud & Falchikov, 2006).

1.2.2 Validity evidence based on internal structure

Internal structure, as a source of validity evidence, relates to the psychometric properties of the instrument and the measurement model used to score and scale assessment items (Downing, 2003). Analyses of the internal structure of a test can indicate the degree to which items that make up an instrument are intercorrelated, that is, appear to be measuring the same construct. A variety of statistical approaches have been used to examine internal structure of an instrument, including factor analysis, item-item and item-total correlations, Cronbach’s alpha, item response theory (IRT) and Rasch analysis (Downing, 2003; J. Hobart & Cano, 2009; Mokkink, et al., 2010; Terwee, et al., 2007; Wilson, 2005; Wolfe & Smith, 2007b).

Issues of bias also pertain to the internal test structure category of validity evidence. Differential item function (DIF) studies can be designed to detect item bias. DIF associated with specific factors or conditions could occur when different groups in a sample, for example, male and female students perform differently on specific items or different clinical instructors assess consistently higher or lower. A useful scale would function consistently irrespective of subgroups within the sample being assessed.

Reliability

Another validation activity that fits within the internal structure category of evidence is the investigation of reliability (Cook & Beckman, 2006; Downing, 2003). Reliability of clinical assessment refers to the extent to which assessment of performance yields relatively consistent outcomes. Reliability has also typically been seen as a quality of an instrument distinct from validity, however Wilson (2005) argues that it is an integral part of validity. An instrument with little or no consistency across different circumstances in which it is applied would be of limited value, no matter how sound other arguments were for its validity (Wilson, 2005).

1.2.3 Validity based on relations to other variables

This source of evidence is primarily based on correlational studies and seeks to provide both confirmatory (convergent), discriminant and predictive evidence of test score validity (Downing, 2003; Streiner & Norman, 2003). The relationship between the results of clinical assessment and other variables hypothesised to measure a construct related to clinical performance provides a form of validation (*convergent evidence*). In addition, clinical assessment scores should not predictably be strongly correlated with scores for an unrelated construct (*discriminant evidence*) (Messick, 1996; Wilson, 2005). Validity might also be inferred if clinical performance scores were good predictors of future performance in the workplace on graduation.

1.2.4 Validity based on response processes

Validity evidence based on response processes is an example of the substantive aspect of validity described by Messick (1989). It is closely related to the previous section (1.2.1) on

content but focuses more on operationalisation of the instrument by examining if assessors are using the instrument as anticipated (American Educational Research Association, 1999; Messick, 1996).

'Think aloud' and exit interviews have been recommended as methods for exploring what educators are thinking during administration of and after completion of an assessment instrument. Data from think aloud interviews may identify aspects of an instrument that are ambiguous or inconsistently interpreted. Additionally, interviews are thought to enable exploration of responses, facilitating better quality feedback than is possible with written response surveys completed at varying times after the assessment has been finalised (American Educational Research Association, 1999; Streiner & Norman, 2003; Wilson, 2005).

Validity evidence based on response processes also includes information on rater training, including content, timing and evaluation (American Educational Research Association, 1999; Beckman, Cook, & Mandrekar, 2005; Cook & Beckman, 2006; Prescott-Clements, van der Vleuten, Schuwirth, Hurst, & Rennie, 2008). Similarly student familiarity with and understanding of instrument format, purpose and score interpretation provides evidence in this category.

1.2.5 Validity based on the consequences of testing (educational impact)

Assessment validity may be indicated by the educational impact of assessment. The Standards (1999) recommend that positive and negative consequences of the testing process be examined and evaluated. This includes the method for determining pass/fail based on scores, any change in student knowledge or skills, and student views on the instrument and its implementation. Provision of formative feedback using an assessment instrument, and the impact of this on student learning, would also contribute to an understanding of the consequences of testing.

1.2.6 Additional sources of validity evidence: acceptability and costs

While not part of traditional validity frameworks, acceptability (which includes time to complete and user satisfaction) and costs of administering performance based assessments

are important features of the complex assessment process necessary to appropriately assess professional competence (Baartman, et al., 2006; Van der Vleuten, 1996a). Investigation of these aspects provides relevant information as assessment processes are not only influenced by educational factors, but also by financial, managerial and institutional values (Baartman, et al., 2007; Streiner & Norman, 2003).

A preliminary search of the literature revealed no systematic review of existing instruments for assessing professional competence of students in physiotherapy programs or any review critically appraising evidence that supports the validity, reliability, acceptability, and impact on education of these instruments. A systematic review was therefore conducted to:

1. Identify all published instruments for assessing clinical performance in physiotherapy practice
2. Summarise the evidence of reliability and validity of each instrument, based on test content, response processes, internal structure, relationship to other variables, consequences of testing (educational impact), acceptability and costs.

1.2 Method

1.2.1 Identification and selection of studies

1.2.1.1 Search strategy

The systematic search was guided by Best Evidence Medical Education (BEME) collaboration guidelines for effective evidence retrieval (Haig & Dozier, 2003b; Harden, Grant, Buckley, & Hart, 1999). Electronic databases were searched without date limits until May 31, 2009 using 'explosions' of key search terms for physiotherapy, assessment, clinical placement, work-based placement, measurement properties, and competency (refer to Appendix 1.1 for an example of a search strategy used to search the CINAHL database). The search strategy utilized search terms derived from the controlled vocabulary used to index articles for each specific database. Truncation and 'wildcard' characters were used to identify variants of words e.g., competen* or competen\$ to locate terms such as competent, competency, competencies or randomi#ed to identify spelling variations.

1.2.1.2 Electronic databases

The databases searched to identify potential studies for inclusion in this review were:

- Index Medicus (Medline via Ovid) 1950 – Present,
- Excerpta Medica Database (Embase via Ovid) 1966 - Present,
- Cumulative Index for Nursing and Allied Health (CINAHL) (via Ovid and EBSCO) 1982 - Present),
- Educational Resource Information Center (ERIC via CSA),
- British Education Index (BEI),
- Allied and Complementary Medicine (AMED),
- PsychINFO,
- Physiotherapy Evidence database (PEDro),
- Cochrane Register of Central Controlled Trials,
- Dissertation Abstracts,
- ISI Web of Knowledge,
- Blackwell Synergy via Wiley Interscience,
- Health and Psychosocial Instruments (HaPI) via Ovid, and
- Best Evidence Medical Education (BEME).

The reference lists of all relevant studies were searched, and follow up searches on the first and second authors of all eligible studies and cited reference searches of eligible articles in the Science Citation Index were conducted. The proceedings of relevant conferences, including The International World Confederation of Physical Therapists (WCPT), The Ottawa International Conference on Clinical Competence and Association for Medical Education in Europe (AMEE) were searched by accessing on-line documents assembled by the sponsoring organisations.

1.2.1.3 Hand searches

Additional hand searches were conducted (for time periods not fully covered by the electronic search) of the journals *Medical Teacher*, *Academic Medicine*, *Medical Education*, *British Educational Research Journal*, *Educational Theory*, *Journal of Educational Measurement*, *Physical Therapy*, *Canadian Journal of Physiotherapy*, and *Physiotherapy*

Research International. Search results were entered into electronic bibliographic management software (EndNote¹) and duplicate records removed. The search commenced in 2007 and was repeated across time with the final search conducted in May 2009.

1.2.1.4 Study selection

To be eligible for inclusion in this review, reports must describe an instrument with the purpose of assessment of professional competence of student physiotherapists in a practice environment and assessment using the instrument must be performed by an educational supervisor (a graduate physiotherapist). The instrument, including the items and scoring system, and data on the application of the instrument in the authentic practice environment (e.g., field testing), must be available for review. Reports relating to the instrument must be available in English or relevant data reported in English. Study participants must be students of any tertiary institute or registration body offering a physiotherapy qualification, re-registration of a qualification or specialisation. Studies were excluded if the instrument was designed for use with standardized patients or used for peer, patient or self assessment but not for assessment of students in the authentic practice environment.

1.2.1.5 Study review

Title and abstract of the studies found through the search procedures were independently screened by two reviewers (JK and MDal). If a report was not able to be deemed ineligible based on title and abstract, full text was obtained. Final determination of eligibility was made after reading the full-text manuscript and evaluation against the predetermined inclusion criteria. Disagreements regarding decisions to include a study were resolved by discussion. Where necessary a third reviewer's opinion was sought (MDav).

1.3 Data extraction

Two review authors (MDal, JK) independently extracted data from all included reports using a standardised form specifically developed to achieve the aims of this review. The reviewers were blinded to each other's data extraction during each stage of the review process. Results from data extraction were compared and discrepancies noted by the two reviewers;

¹Endnote X2, Thomson ResearchSoft, www.thomsonresearchsoft.com

if necessary a third reviewer (MDav) repeated the data extraction and the final set of extracted data represents the agreed decisions of 2 or 3 reviewers.

Characteristics of the identified instruments were extracted and included, domains of practice, items, rating scale development, scoring criteria and compilation and interpretation of total instrument score (Table 1.1). In addition data was extracted within each of the five categories of validity evidence recommended by the Standards (1999). Questions to assess validity evidence in the 5 categories using best practice criteria proposed in the literature were developed and are presented in Table 1.2 (American Educational Research Association, 1999; Baartman, et al., 2006, 2007; Streiner & Norman, 2003; Terwee, et al., 2007; Wilson, 2005; Wolfe & Smith, 2007b; Wolfe & Smith, 2007a). Each instrument was assessed against these categories of validity evidence enabling identification of what is known about the quality of the instruments for assessing professional competence.

As recommended by Terwee et al (2007) the quality assessment criteria (Table 1.2) were not summarised into one overall quality score. Summing the scores for each criterion into one overall score would assume all quality criteria are equally important which may not be the case. Data were therefore synthesised descriptively and presented for reader consideration.

Due to the variable instruments and methods employed to develop and test instruments in the included studies, meta-analysis was neither possible nor logical.

Table 1.1: Characteristics of the instruments identified in the review

	Instrument name, country, source(s)	Domains and items	Response Options and Scoring
1	Clinical performance instrument (CPI) for Physical Therapists, USA and Canada Task force 2002 (Task Force for the Development of Student Clinical Performance Instruments, 2002) (Straube & Campbell, 2003) (Tsuda, Low, & Vlad, 2007) (Adams, Glavin, Hutchins, Lee, & Zimmerman, 2008) (Logemann, 2006)	24 items and 2 global performance items	Visual analogue scale 0-100mm (0 = novice, 100 = entry level). Additional 'with distinction' 'significant concerns' & 'not observed' checkboxes. The rater must consider 5 performance dimensions during the assessment process: quality of care, supervision/guidance required, consistency of performance, complexity of tasks/environment & efficiency of performance. Total possible raw score 0 to 2400. Total final score is the average of the item scores: range 0 to 100.
2	Evaluation of Clinical Competence (ECC), Canada (Loomis, 1985a, 1985b). (P. D. Cox, La, & Pappachan, 1999)	2 domains, 40 items Patient care (31) Professional behaviour (9)	Patient care: 5 point rating scale (0 – 4) ranging from 1 = incompetent to 4 = highly competent. 0 = not observed . Professional behaviour: 5 point rating scale (0 – 4) based on the frequency of the behaviour. 1= inconsistently to 4 = always. 0 = not observed. Each item weighted 1 – 3 (1 = important, 2 = ?, 3 = essential). Possible weighted score range 0 to 380. Total score for each section obtained by summing weighted score on each item. No detail provided on which competencies weighted and by how much.
3	Clinical Internship Evaluation Tool (CIET), USA (Fitzgerald, Delitto, & Irrgang, 2007)	2 domains, 42 items Professional behaviours(18) Patient management skills (24)	5-point rating scales for the items and a global rating scale. Professional behaviour (0 = never displays the behaviour, to 4 = always displays behaviour). Student must score at least 4 for all items. Patient management (1 = performance well below competent clinician, to 5 = well above competent clinician. Student must score at least 3 for all items. Global rating scale 0-10 (5 = at the level of a competent clinician) Item scores are summed to possible total score 0 to 120
4	Un-named instrument, Ireland and UK (Meldrum, et al., 2008) Author reports the assessment form is based on the work of (Cross, 2001) with modifications.	3 domains, 36 items Patient management (15), Professional development (15), Organisation and management (6)	Patient management 0 to 600, Professional development 0 to 300 Organisation and management 0 to 100 Grading system: 0-49 = fail; 50-59 = third class honours; 60-64 = 2 nd class honours grade 1 (2.1); and 70-100 = 1 st class honours. Grading guidelines not provided

	Instrument name, country, source(s)	Domains and items	Response Options and Scoring
5	Common Assessment Form (CAF), Ireland (Coote, et al., 2007) (based on the work of Meldrum 2008)	5 domains, 40 items Patient assessment(10), Patient treatment (10), Professionalism (10), Documentation (5), Communication (5)	Each item scored 0 to 10 First 3 domains each scored 0 to 100 Last 2 domains each scored 0 to 50 Possible total score 0 to 400 Student pass = score > 50/100
6	Physical Therapist Manual for the Assessment of Clinical Skills (PT MACS) (Stickley, 2002, 2005) (Logemann, 2006)	PT MACS: 42 items with 11 additional situational items Maximum possible items = 53 Note: PT MACS superseded the B MACS	5 point rating scale from Below expectations to Above Expectations. If item not rated, item has not been observed Academic coordinators of clinical education or directors of clinical education decide student grade based on the PT MACS and feedback provided by the Clinical instructor.
7	Mastery and assessment of clinical skills (Blue (B) MACS), USA (Hrachovy, et al., 2000)	Domains not defined B MACS: 38 items with 12 additional situational items.	4 point rating scale from Does not meet entry level standard to Exceeds entry level standard. No information on possible total score range.
8	Student clinical competence scale (SCCS), USA (Rheault & Coulson, 1991)	6 items Knowledge base, Communication skills, Safe treatment techniques, Effective treatment techniques, Problem solving, Professionalism	4 point scale: 0 = poor, 1 = low average, 3 = high average & good. Possible total score 0 to 18.

Table 1.2: Summary of evidence for validity, reliability, acceptability and costs

Validity Evidence based on	Criteria	1.CPI	2.ECC	3.Meldrum	4.CAF	5.CIET	6.PT MACS	7.B MACS	8.SCCS
Test content	Were characteristics of study participants reported?	✓	✓	✓	✓	✓	✓	✓	✓
	Were personnel involved in the instrument development specified?	✓	✓	✓	✓	✓			✓
	Were students involved in the development of the instrument?	✓							
	Was a pool of items generated?	✓	✓	✓	✓	✓			
	Were the criteria for item pool reduction to the final item list specified?		✓						
	Were the criteria for technical quality of item design outlined?								
	Were performance indicators included?	✓	✓	✓	✓		✓	✓	
	Was the final item set mapped against relevant standards?						✓		
	Was the process of development of the rating scale reported?								
	Was the rating scale described: format, width, descriptors and scoring criteria?	✓	✓	✓	✓	✓	✓	✓	✓
	Was the ability of the rating scale to differentiate levels of competence investigated?								
Was information provided on compilation and interpretation of total score?	✓		✓	✓	✓				
Was the instrument tested in an authentic clinical environment?	✓	✓	✓	✓	✓	✓	✓	✓	
Internal structure	Was factor analysis performed on an adequate sample size? (7 * # items and >100) ¥	✓				✓	✓		
	Was Cronbach's alpha calculated and between 0.70 and 0.95? ¥	✓				✓			
Reliability	Was any bias for items among subgroups in the sample investigated eg DIF?								
	Was any investigation using IRT or Rasch analysis conducted?	✓					✓		✓
	Was an inter-rater reliability trial conducted?	✓	✓	✓	✓				
	If reliability studied, was the number of raters specified?			✓					
	If reliability studied, was the number of students assessed specified?		✓						
	If reliability studied, was the number of paired assessments provided?	✓	✓	✓	✓				
	If reliability studied, were raters blinded to other rater test scores?	✓	✓	✓	✓				
	If reliability studied, were Test 1-2 mean score differences (<i>d</i>) & SD of diff (<i>sd</i>) or comparable data reported?	✓	✓	✓	✓				
If reliability studied, was the test1 /test 2 correlation specified?	✓	✓	✓	✓					
Response Processes	Were interviews of assessors and test takers (students) performed?								
	Were details of the content of training reported?								
Relationship to other variables	Were relationships to other tests hypothesized to measure related or different constructs analysed?.	✓	✓		✓	✓			
Consequences (educational Impact)	Was the instrument used to provide feedback to the students?	✓	✓			✓	✓	✓	
	Was student learning evaluated?								
	Was feedback on instrument use sought from students?	✓							
Acceptability	Was the acceptability of instrument to stakeholders formally investigated?	✓	✓			✓		✓	
	Was the time taken to complete the instrument reported?					✓		✓	
Costs	Was there any information provided on costs related to instrument use?								

✓ = Yes, criteria addressed; No tick = criteria not addressed or there is insufficient information to decide; * = multiply, # = number, DIF = differential item functioning, IRT = item response theory, ¥ = values chosen based on recommendations from Terwee et al (2007).

1.4 Results

1.4.1 Flow of papers through the review

Database searches retrieved 1364 papers. Hand searching and searching of remaining sources identified 14 additional potentially relevant articles. After screening title and abstract and eliminating duplicates, 46 papers remained for detailed analysis. Evaluating full text against the inclusion criteria, 31 were eliminated. A total of 15 studies reporting on eight instruments met the inclusion criteria (Figure 1.2).

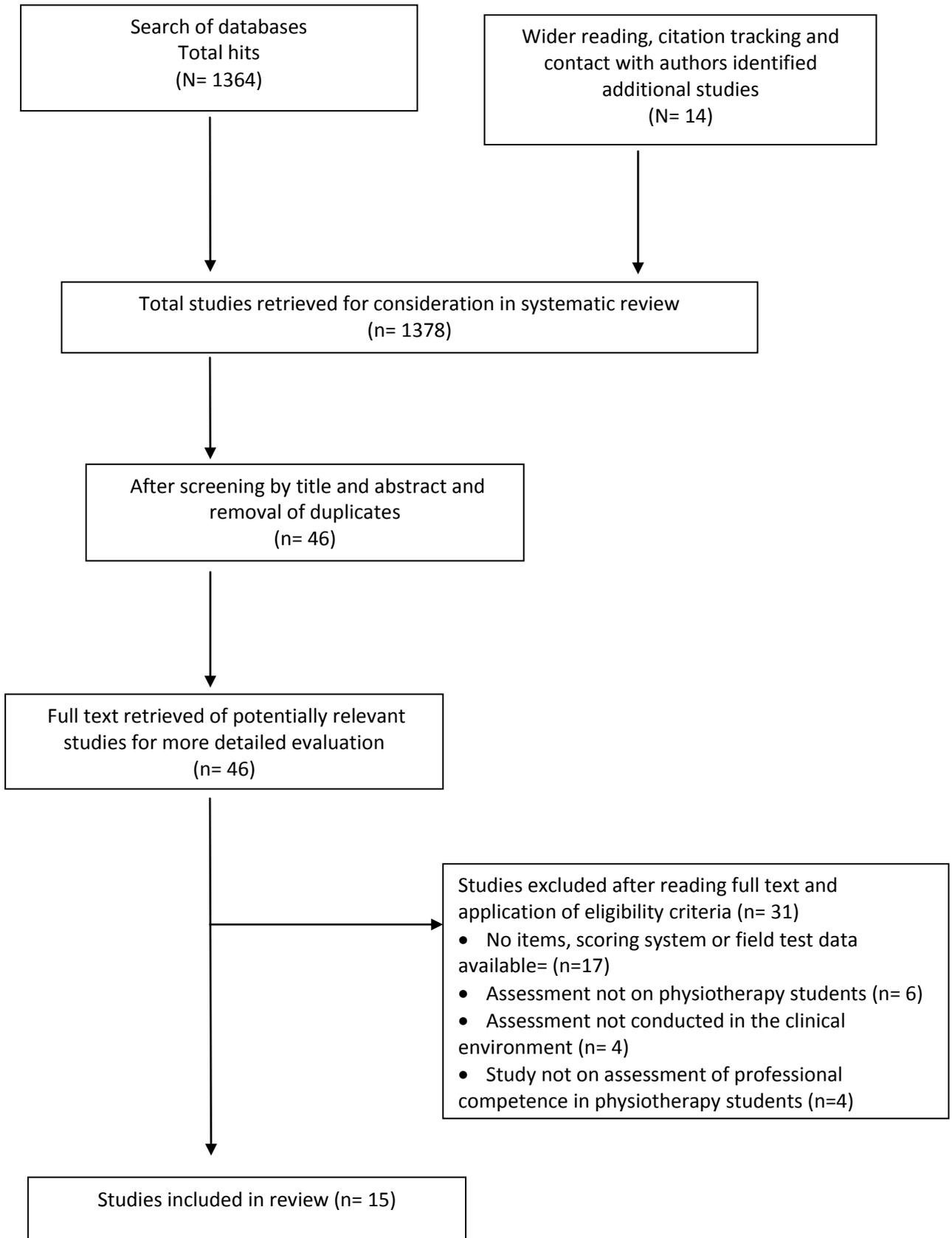


Figure 1.2: Flow of papers through the review

1.4.2 Description of instruments

The characteristics of each instrument are presented in Table 1.1. Of the eight instruments identified all contained items covering the two primary domains of professional practice and patient management. The number of items varied from six in the Student clinical competence scale (SCCS) developed by Rheault and Coulson (1991) to 53 in the Physical Therapist Manual for the Assessment of Clinical Skills (PT MACS) developed by Stickley (2002). Scoring criteria and type of rating scale was varied. A 100mm Visual Analog Scale (VAS) was used in the Clinical Performance Instrument (CPI) developed for use in the USA and Canada (2002), while other instruments used 4 point (Blue MACS, SCCS), 5 point (PTMACS, Clinical Internship Evaluation Tool (CIET), Evaluation of Clinical Competence (ECC)) and 10 point rating scales (Clinical Assessment Form (CAF)).

Table 1.2 provides a summary of evidence of validity based on test content, response processes, internal structure (including reliability), relationship to other variables, consequences of testing (educational impact), acceptability and costs (complete data for each instrument is presented in Appendix 1.2). Study participants were undergraduate or graduate entry students in all studies. No instrument was located that was used to assess re-registration of a qualification or post graduate specialization.

Five of six authors contacted for additional information replied: three provided copies of assessment instruments (Coote, et al., 2007; Cross, 2001; Meldrum, et al., 2008), one provided abstracts from conference presentations (Adams, et al., 2008) and one confirmed that no reliability data had been published (Fitzgerald, et al., 2007). The sixth author could not be located, so the study was excluded as the instrument was not available for review (Morris, 2006).

1.4.3 Validity evidence based on test content

Tables 1.1 and 1.2 reveal that assessment instruments were typically developed so that a number of domains (eg professional behaviour, practical skills) were assessed. Within each domain, one or more items were scored. Despite the variation in the total number of items

(6 to 53) and domains (2 to 11), there was considerable homogeneity in instrument content. This reflects concordance regarding the expected competencies of graduate physiotherapists worldwide. For five of the eight instruments (63%), evidence was provided in relation to validity based on test content with respect to measurement aim, target population, item selection, personnel involved in instrument development and format of the rating scale. The technical quality of the items is also recognised as a source of content-related validity (Downing, 2003; Streiner & Norman, 2003; Wolfe & Smith, 2007a). None of the studies presented any data on the criteria used to ensure that technical quality of the items was achieved.

Most studies reported consideration of relevant professional standards in development of the item pool although a structured mapping exercise was reported for only the PT MACS (Stickley, 2002, 2005). Performance indicators, intended to serve as a learning guide for students and provide educators with examples of behaviours that would indicate competence for each item, were included as part of six of the eight (75%) instruments.

While all instruments reported the scoring criteria and the type of rating scale utilised, no study described the process for development of the rating scale. In addition, little information was provided on the compilation, interpretation and application of the total score for four of eight instruments (50%). No report included information on how a pass standard was determined, or how the rating scale functioned to differentiate between levels of professional competence. In addition, involvement of students in instrument development occurred in only one study (13%). All eight instruments were tested in the authentic practice environment.

Overall for 6 of the 8 (75%) instruments (CPI, ECC, CIET, CAF, PT MACS and the unnamed instrument developed by Meldrum and colleagues), there was sufficient information to enable consideration of validity evidence other than that based on test content (refer to Table 2) (Coote, et al., 2007; Fitzgerald, et al., 2007; Logemann, 2006; Loomis, 1985a, 1985b; Meldrum, et al., 2008; Stickley, 2002, 2005; Task Force for the Development of Student Clinical Performance Instruments, 2002).

1.4.4 Validity evidence based on response processes

No study reported conducting think aloud interviews while an educator or student completed an assessment instrument or exit interviews following completion of the instrument. All studies reported that raters received some training in the use of the instrument but the actual content, delivery method and evaluation of training was not reported.

1.4.5 Validity evidence based on internal structure

Evidence of internal consistency was found for three instruments (CPI, CIET, PT MACS). Results of both factor analysis and Cronbach's alpha were provided for two instruments (CPI and CIET). Logemann (2006) investigated the internal structure of the CPI and PT MACS using principal components analysis (PCA) and item response theory (IRT) using a 2-parameter graded response model. Applying a principal components analysis the Task Force for the Development of Student Clinical Performance Instruments (2002) found two factors (physical therapy specific clinical skills and professional behaviour) accounted for 12% and 9% respectively of observed variance in CPI scores. Logemann (2006) also found two similar factors accounting for 73.8% (66.6% and 7.2%) of the variance in the CPI and two factors explaining 54% (42.7%, and 11.2%) of the variance in the PT MACS. Employing a similar approach, Adams et al (2008) found three factors explaining 72% of the variance in CPI scores. These were labeled integrated patient management (54.2%), professional practice (12.2%), and career responsibilities (includes resource and fiscal management and career development/lifelong learning), (5.4%). Cronbach's alpha for CPI item scores ranged from 0.96-0.97 for the whole instrument and from 0.75-0.96 for item scores in the three domain subscales (Adams, et al., 2008; Task Force for the Development of Student Clinical Performance Instruments, 2002). The CIET, from its inception, was considered to be assessing two distinct domains of practice, professional behaviour and patient management (Fitzgerald, et al., 2007). Factor analysis, Cronbach's alpha and item-to-total scale score correlations were determined separately for these two domains. The professional behaviour domain did not conform to a one factor model (3 factors emerged from 18 professional behaviour items), so Cronbach's alpha and item-to-total correlations were not calculated. In the patient management domain only one distinct factor was

extracted. Cronbach's alpha averaged 0.98 when calculated using data collected across seven clinical blocks. Logemann (2006) collapsed the 100mm VAS that is used to score CPI items into the six categories recommended by Straube and Campbell (2003) prior to analysis. Logemann's results therefore do not indicate the properties of the CPI as it appears to be currently used. Assessment of possible bias in items (differential item function (DIF)) was not investigated for any of the eight instruments.

Reliability

Authors reported Intraclass Correlation Coefficients ICCs (point estimates only) for inter-rater reliability of four instruments, CPI (.87), Meldrum et al (2008) (.84), CAF (.84) and ECC (.62). These studies were conducted in the clinical environment on completion of standard clinical placement blocks of four to six weeks duration, and were reported by the authors to indicate an acceptable level of inter rater reliability. The number of pairs of raters in all studies ranged from 35 to 86 however the number of students and clinical educators involved in each trial was generally not provided.

The ICC, the Standard Error of Measurement (SEM) and the Minimal Detectable Change at 90% confidence interval (MDC_{90}) can all be derived from analysis of item scores when assessment is repeated under circumstances when no real change in the underlying construct of interest is expected. Each index conveys distinctly different information, and all indices are relevant to a view of measurement reliability. In the case of measurements of competency to practice, the ICC provides information on the consistency with which scores enable ranking of students, the SEM provides information on the magnitude of error associated with a single assessment score expressed in the units of measurement, while the MDC_{90} describes the magnitude of change required for confidence that (in 90% of cases) real change in the underlying construct has occurred. Only two studies reported data that enabled calculation of the SEM and MDC_{90} . For the Common Assessment Form (CAF) developed by Coote et al (2007) SEM and MDC_{90} were 4% and 8.9% of a 0-100 point scale. For the instrument developed by Meldrum et al (2008) (also a 0-100 point scale) these were 2.1% and 4.8% respectively. These statistics were computed by the author (MDal) using data reported by Coote et al (2007) and Meldrum et al.(2008). There was insufficient data provided in reports of other inter-rater reliability studies to enable calculation of the MDC_{90} or SEM.

1.4.6 Validity evidence in relation to other variables

Four of the eight instrument developers analysed the relationship between the results of clinical assessment and other measures hypothesised to measure a related or different construct to clinical performance. The Task Force for the Development of Student Clinical Performance Instruments (2002) investigated the relationship between CPI scores and social competence measured with the social skills inventory (SSI) (n=31) to assess whether the CPI measured social skills or the clinical performance of the students. The authors reported that the Pearson correlations between the SSI total score and CPI item scores were low and not significant (range .02 - .25). Loomis (1985a) correlated total ECC scores with hiring ratings for fourth year students, where educators were asked to rate their willingness to employ the student. Although the sample was small (n=25), the two ratings were significantly correlated (Spearman's rho 0.68, p=.001). In examining the mean item scores and total instrument scores for different clinical blocks, the CPI, ECC and CIET demonstrated significant improvements across time supporting the hypothesis that the instruments displayed the expected improvement in scores with clinical experience (Fitzgerald, et al., 2007; Loomis, 1985a; Task Force for the Development of Student Clinical Performance Instruments, 2002).

In the study of the CPI, data from a subset of students (n=68) who were on either their first or final clinical block, were used to perform a known groups analysis using a t-test for differences between mean scores (Task Force for the Development of Student Clinical Performance Instruments, 2002). Nineteen of the 24 items demonstrated differences between student scores on their first and final clinical units ($p = .0002$). The six items not demonstrating difference were varied and included item two (presents self in a professional manner), item 18 (addresses patient needs for services other than physical therapy) and item 24 (addresses prevention, wellness and health promotion needs of individuals, groups and communities).

The rate of improvement in student performance across time measured by the CPI and CIET was not linear. Items representing aspects of patient management continued to improve from first to last clinical placement, while mean scores for other items (e.g. presents self in a professional manner) did not change significantly across time. This item was hypothesised to be an aspect of

clinical performance mastered early in a student's professional practice program and thus not expected to continue to improve with training.

CPI scores for six items assessing patient assessment and treatment correlated strongly with the number of clinical days experience (Pearson correlations ranged from .34 ($p=.0001$) to .40 ($p=.0001$)). For CIET scores (Fitzgerald, et al., 2007), the correlations between patient management scores and a global rating score for overall clinical performance for seven clinical blocks, ranged from 0.54 to 0.89 (Spearman Rho, $p= .01$).

Coote et al (2007) compared the test results from their newly developed instrument (CAF) with scores for existing instruments from four Irish Universities offering physiotherapy programs (Royal College of Surgeons in Ireland, Trinity College Dublin, University College Dublin and University of Limerick). The validity/reliability of four comparison instruments is unknown, but high correlations were found with all four instruments as demonstrated by Pearson correlation coefficients ranging from 0.88 for the University College Dublin instrument to 0.98 for the University of Limerick (UL) instrument. The UL assessment instrument was the form most similar to the CAF.

The relationship between total scores on the CPI and results on the written National Physical Therapy Examination (NPTE) were examined (Task Force for the Development of Student Clinical Performance Instruments, 2002). Logistic regression (for data from 126 students) demonstrated that neither a CPI subscale score nor the total score was associated with NPTE results. A student's cumulative Grade Point Average (GPA) however, correctly classified 97% of those who passed the NPTE on first sitting.

1.4.7 Validity evidence based on consequences of testing (educational impact)

No study examined the effect of the assessment process on student learning, transfer of skills to new situations or change in the quality of patient care. One study, Task Force for the Development of Student Clinical Performance Instruments (2002) investigated student perceived satisfaction with the assessment instrument; however the surveys focused on procedural aspects of instrument use rather than aspects of student learning. Five of the eight instruments were used to provide formative feedback to students on their

performance during the clinical placement but data on the perceived effect of this was not reported.

1.4.8 Additional sources of validity evidence: acceptability and costs

Data on user satisfaction were reported for four instruments (CPI, CIET, ECC and B MACS) via questionnaires and/or focus groups (Fitzgerald, et al., 2007; Hrachovy, et al., 2000; Loomis, 1985a, 1985b; Task Force for the Development of Student Clinical Performance Instruments, 2002). Clinical educators who chose to complete the questionnaire reported being satisfied with these instruments. On average, the time taken to complete the CIET was 30-60 minutes and the B MACS 1.6 hours. No data were found regarding the time taken to complete the other six instruments.

All of the eight instruments were tested by clinical educators longitudinally assessing students within an authentic clinical environment following completion of clinical placement blocks ranging in length from 2- 12 weeks. No data were found on any aspect of the costs associated with the development and evaluation of the instrument or of the costs of assessing clinical performance to the facilities supervising students.

1.5 Discussion

The review found eight instruments developed to assess the professional competence of physiotherapy students in the clinical environment.

Validity evidence based on test content was sufficient for most instruments. All studies engaged with a broad cross section of experts and relevant stakeholders during development of the instruments. This enabled comprehensive item content and appropriate representation of the construct (professional competence) being measured was probably achieved. In the majority of studies (75%) the quality of the final item set was developed and refined through iterative review procedures, for example, focus groups and pilot trials of the instrument on the target population. However none of the studies presented any data on the criteria used to ensure that technical quality of the items was achieved and only one instrument, the CPI, involved students in instrument development and refinement.

While the relevant professional standards guided the initial development of the item content for each instrument, a formal mapping of the instrument content to develop alignment with the relevant standards was conducted only for the PT MACS (Stickley, 2005). Inherent in effective and accurate use of any measurement instrument is information on the compilation and interpretation of the total score. No information was provided for the SCCS or the B MACS (Hrachovy, et al., 2000; Rheault & Coulson, 1991). A further two instruments, the ECC and PT MACS (Loomis, 1985b; Stickley, 2005), provided insufficient information to enable compilation and interpretation of the total score. For example, the total score on the ECC was obtained by summing the weighted score on each item, but which items were weighted and how they were weighted was not described.

A central issue surrounding the scoring method of the instruments was the level of measurement provided by the method. As a student's ability increased, so did their score on the rating scales. Several authors (Coote, et al., 2007; Fitzgerald, et al., 2007; Loomis, 1985a; Meldrum, et al., 2008; Task Force for the Development of Student Clinical Performance Instruments, 2002) use the method of summated ratings as proposed by Likert (1952). In this method, consecutively ordered response options are allocated sequential numbers with the item scores being summed to give a total score. Scores for items are treated as interval level data. While a variety of scoring systems have been used, from 100mm VAS to 4, 5 and 6 point behaviourally anchored scales, all scoring systems provided ordinal rather than interval level data and the validity of parametric analyses was not established.

The internal structure of three instruments was investigated using factor analysis and Cronbach's alpha. The CPI, PT MACS and CIET appear to be multidimensional instruments measuring at least two constructs: physiotherapy specific clinical skills and professional behaviour. According to the Rasch Measurement Model (RMM) the sum of item scores is difficult to interpret in the context of multidimensional scales (Bond & Fox, 2007). To illustrate, if a scale contained items that either assessed professionalism or assessed practical skills in equal measure, a student scoring 75% might have excellent professionalism and weak practical skills or vice versa. Two students with the same assessment outcome might, in this situation, have very different skill sets. One way to overcome this problem is to identify the items that assess particular domains, and score the domains separately. Further research into the measurement properties of assessment instruments using Rasch

analysis would assist to clarify this issue of dimensionality, remove sample and scale dependency and provide a validated approach to total score collation and interpretation.

Of importance in relation to the internal consistency of an instrument is the absence of any item bias among subgroups in the sample investigated. There were no data presented for any of the eight instruments regarding testing for item bias. This is a noticeable deficit in the validity evidence provided in support of the eight instruments identified in this review.

When considering intraclass correlation coefficients, Streiner and Norman (2003) recommend that a coefficient of 0.75 and above is a minimal requirement for a useful instrument. Landis and Koch (1977) similarly recommended that coefficients between 0.61 and 0.80 represented substantial strength of agreement between measurements, whereas Portney and Watkins (1993) recommended 0.90 for making decision about individual subject scores. Studies identified in this review demonstrated intraclass correlations (2,1) ranging from .87 for the total CPI score (Task Force for the Development of Student Clinical Performance Instruments, 2002) to 0.84 for the two Irish instruments (Coote, et al., 2007; Meldrum, et al., 2008). Loomis (1985a) found ICC's of 0.62 and 0.59 for third and fourth year total scores respectively. Overall this in all probability represents acceptable levels of inter-rater reliability for these instruments. The ICCs provide information on the utility of measurements to differentiate between scores for different individuals and are an index of consistency in ranking order. However, what is lacking in this approach is that it does not provide information about the magnitude of error (expressed in the scale units of measurement) associated with a single application of the test, or repeated applications of the test under conditions when it is reasonable to expect that the underlying construct has not changed (Streiner & Norman, 2003; Weir, 2005).

The Standard Error of Measurement (SEM) and MDC_{90} could be estimated for only two instruments. The SEM for the total score on the CAF was 4.0 points and 3.0 for the instrument developed by Meldrum et al (2008), (scale width 0 – 100).

The 95% confidence band around a single score was calculated by the author (MDal) using

Equation 1.

$SEM \times t$

Equation 1

Where;

t = the appropriate t value at a df = (n (number of pairs) -1), at an alpha level of 0.1.

The 95% confidence band around a single score was 8 CAF points (given $t(0.05, df= 70) = 1.99$) and 6 points (given $t(0.05, df= 85) = 1.98$) on the instrument developed by Meldrum.

The SEM has implications for students whose score is within the borderline pass/fail range. If the pass mark is 50 out of the total 100 marks, then 50 minus a SEM of 8 marks on the CAF (42) might be considered an outright fail, while 50 plus 8 marks (58) might be considered an outright pass. The values in between would require a process for deciding on further assessment for confidence that the student has an adequate level of professional competence on the items for which scores are poor. As none of the 15 studies provided SEM estimates, no study discussed identification and management of students whose clinical performance was scored as borderline pass/fail.

Overall there is limited published data supporting the reliability of four instruments identified in this review (Coote, et al., 2007; Loomis, 1985a; Meldrum, et al., 2008; Task Force for the Development of Student Clinical Performance Instruments, 2002).

Because there are no gold standard instruments for assessing professional competence in physiotherapy students, correlating data from a newly developed instrument and an existing assessment instrument is of unknown value (Prescott-Clements, et al., 2008; Smith, 2001). Similarly, testing the predictive validity of an instrument by examining the relationship between instrument scores and results on a national physical therapy written examination (NPTE) may also be inappropriate. Lack of correlation between these results may only signal limited constructive alignment between the skills required to pass the NPTE and those essential for adequate entry level clinical performance.

An important indicator of the acceptability of an instrument in a busy clinical environment is the time it takes the clinical educator to complete testing. This aspect was not reported for 6 of the 8 instruments. What clinical instructors consider an acceptable time to complete an

assessment instrument is not known, nor is there sufficient information about the costs involved in the assessment of clinical performance or other aspects of the assessment process that clinical instructors found satisfactory or unsatisfactory. A potential level of bias in the findings relating to user satisfaction could also be present since only those clinical instructors who self selected to return their questionnaires were included in reported results.

Data examining whether educators were interpreting and using the items, performance indicators and response scale as intended has not been reported for any instrument, nor is it clear if aspects of instruments were ambiguous or inconsistently interpreted. In alignment with the findings of Pelgrim et al (2010), this review found that all studies commented on the importance of training, however none provided sufficient information regarding content and delivery. Additionally, data evaluating the influence of training on instrument use or on inter-rater reliability results was not presented and represents an area requiring future research.

Most instruments were used to provide feedback to students during and at the end of the clinical placement. This is representative of the importance placed on the role of formative feedback by the researchers and clinical educators but no research has investigated if students find feedback based on the instrument of benefit or if feedback facilitates a change in performance. Similar findings have been found in reviews examining work-based assessment of medical students (Kogan, Holmboe, & Hauer, 2009; Pelgrim, et al., 2010).

Research concerning instrument development and evaluation in the area of assessment of clinical performance has evolved in the last ten years and existing research is yet to incorporate recommendations regarding contemporary best practice in instrument development. In particular, utilising Item Response Theory or Rasch analysis as the measurement or statistical model during development and/or evaluation of an instrument has been recommended but not employed (J. Hobart & Cano, 2009; McAllister, Lincoln, Ferguson, & McAllister, 2010; Streiner & Norman, 2003; Wolfe & Smith, 2007a). The choice of measurement model used affects the type and quality of data available as evidence of validity supporting the proposed use of the assessment instrument. A critical advantage of

Rasch measurement is that it enables the abstraction of equal units of measurement from the raw data of observations, i.e. scores on the items of an assessment tool. These can be calibrated and then used with confidence to measure and quantify human attributes such as competence in physiotherapy practice (Bond & Fox, 2007). This conversion facilitates appropriate interpretation of differences between individuals and allows tallying of scores to provide a meaningful total test score. In addition, Rasch analysis enables testing of the internal construct validity of a scale for unidimensionality (considered an essential quality of a scale scored by adding results of items) and identification of gaps in the targeting of items to the students' abilities. Rasch analysis also enables assessment of item bias through investigation of differential item functioning.

This review helped to define the criteria essential to consider when developing and evaluating an assessment instrument. The criteria were often absent or difficult to locate without personal contact with authors. Use of guidelines, such as Preferred Reporting Items for Systematic Reviews and Meta- Analyses (PRISMA), when structuring a report would enable comprehensive reporting of all relevant data, replication of studies, comparison of results and facilitate building on previous research ensuring the continued evolution of performance based assessment (Moher, Liberati, Tetzlaff, & Altman, 2009).

1.5.1 Implications for practice

The review found eight instruments developed to assess professional competence of physiotherapy students within the clinical environment. The review failed to identify convincing evidence sufficient to support one instrument above others. The eight instruments differ in number of items, type of rating scale, and scoring criteria. Overall, the CPI, ECC, CIET, CAF and one unnamed Irish instrument are currently supported by the most evidence in relation to the development, refinement and evaluation of their measurement properties, though no instrument has been comprehensively investigated.

These instruments have been developed by and for the American, Canadian and Irish physiotherapy professions. As such they may or may not be appropriate for use within the physiotherapy profession of Australia due to differences in professional standards as defined by the Australian Physiotherapy Council (2006).

Since the review did not locate any study investigating instruments used to assess re-registration of a qualification or post graduate specialization, currently there is inadequate published evidence of validity and reliability of the assessment processes for these participant groups. Attention to the systematic collection of validity evidence for scores from assessment instruments will improve the outcomes for all stakeholders in physiotherapy education.

1.5.2 Limitations of the review

A limitation of this review was that articles were only included if they were in English. Additionally, there was no standard list of criteria essential for best practice in development and evaluation of instruments to assess professional competence in the workplace of students of the health professions. Assessment of the measurement properties of the instruments located in this review required collation of criteria from several sources, many of which related to instruments designed to measure health outcomes and quality of life. As such, the list of criteria outlined in Table 1.2 is both novel and unverified in this context.

2. Chapter Two: The design and development of an instrument to assess professional competence in physiotherapy students

2.1 Introduction

The development of an instrument to assess the performance of a student in the clinical context requires considerable investment of both human and fiscal resources and should be embarked upon with consideration of the quality of existing instruments (Streiner & Norman, 2003). The systematic review (Chapter One) found eight instruments that have been developed to assess performance of physiotherapy students within the clinical environment. No single instrument was supported by comprehensive evidence of validity and utility. In addition, there was insufficient evidence to support choosing one instrument above others for use in the Australian and New Zealand context where physiotherapists have roles, scope of practice and registration requirements that have been influenced by the specific demands of accrediting bodies and the Australian healthcare system.

The systematic review highlighted that there is considerable consensus in assessment instruments that have been developed with respect to the domains of practice typically assessed (professional behaviour and practical skills). It also indicated the need to develop an instrument that is mapped to Australian and New Zealand (ANZ) standards.

In 2006, when this research commenced, there were 13 entry level physiotherapy programs in Australia. In 2010, when this thesis was completed, there were 16 programs with a further two programs in the planning stages. For graduates to be eligible for registration to practice as a physiotherapist on graduation, programs are required to be accredited by the Australian Physiotherapy Council (APC), formerly the Australian Council of Physiotherapy Regulating Authorities (ACOPRA). The Australian Standards for Physiotherapy provide the blueprint that all programs follow to ensure that a minimum set of performance standards is achieved by graduates.

Despite this national accreditation process and single set of performance standards, each Australian and New Zealand physiotherapy program (in 2006) had developed its own instrument to assess student performance in the clinical setting. These instruments evolved

from the need for programs to establish a method for assessment of clinical competence in the absence of an accepted standard. Variation in assessment practices reflected the proposal by Newble et al. (1994) that educational methods are commonly based on historical practices and personal experiences of assessors, rather than on evidence obtained through rigorous research of instruments and their measurement properties. As highlighted by the systematic review, no validity or reliability data were found for any of the instruments in use in ANZ. For clinical educators accepting students from multiple programs, the diversity of assessment forms and supporting documentation represented a substantial burden. As new physiotherapy programs commenced, this burden multiplied.

In 2006, with the support of all universities in Australia and in New Zealand, this research commenced with the aim of developing a single national assessment instrument aligned to the needs of local users and accreditation standards. The test development was supported by a grant from the Carrick Institute for Learning and Teaching in Higher Education Ltd (now the Australian Learning and Teaching Council). The process of instrument development was planned with consideration of:

- validity of the measurements
- reliability of assessment outcomes
- refinement of the instrument utilising feedback from educators and students
- alignment of the instrument with best practice in assessment
- feasibility of the instrument for monitoring and measuring performance in the practice environment
- utility of the instrument for educators and students. For educators the instrument needed to provide a vehicle for valid assessment of performance and suitable formative feedback to guide the development of desirable performance. For students the instrument needed to facilitate appropriate reflection and describe unambiguous performance targets.

2.1.1 Developing instruments to measure performance in the clinical context.

Professional education programs typically assess across all levels of clinical competence and include direct assessment of clinical practice. The assessment of adequate performance is required for certifying fitness to practice. Professional practice necessitates understanding and dealing with highly variable circumstances and assessment is therefore difficult to standardise across students (Rethans, et al., 2002). A proposed solution to this complexity is to monitor students over a sufficiently long period of time to enable observation of practice in a range of circumstances and across a spectrum of patient types and needs. This has been argued as superior to one off 'exit style' examinations (van der Vleuten, 2000). Longitudinal context based assessment also enables assessment to encompass local contexts, cultures and workplaces within which learners must demonstrate competence and adaptability; it enables the important sociocultural perspective of learning to be addressed. Students are able to construct their own learning within the context specific clinical environment and develop the essential professional habits of reflective practice and ongoing learning (Higgs & Bithell, 2001; Sfard, 1998). Longitudinal assessment can be subtle and continuous or intermittent and structured and is recognised as the preferred approach to gather a reliable and valid representation of students' skills in workplace practice.

Longitudinal assessment of professional competence of physiotherapy students in the workplace is the assessment approach used within all ANZ physiotherapy programs. Clinical educators (registered physiotherapists) generally rate a student's performance on a set of items on completion of a block of workplace practice which may range from 2-15 weeks. Physiotherapists assist people to regain physical function and health, and this frequently involves support for people with substantial impairment. The work requires judgments about what activities are sensible and safe to attempt, and students are usually supervised closely in 1:1 up to 1:4 supervisor student ratios. Educators typically expose students to increasingly challenging tasks as they gain experience, and occasions of close supervision characterise student placements just prior to graduation. Students complete up to 42 weeks of clinical education, providing opportunities for repeated assessment of professional competence on multiple occasions across a spectrum of required clinical areas and workplaces.

ANZ instruments available at the time this project began were designed to map clinical performance on a continuum, with each student's ability varying from less to more. Evaluations of the student's ability were made based on the analysis of numbers generated when students were rated on each item. If correct interpretations of such scores are to be made, the instrument must be both psychometrically sound and educationally informative (Prescott-Clements, et al., 2008; Streiner & Norman, 2003). These requirements were fundamental considerations in the development and evaluation of a new assessment instrument.

This Chapter presents the theories relevant to instrument development that underpin the design of the instrument development processes, while Chapter Three describes the process taken in creating the first version of an instrument suitable for assessing professional competence in ANZ physiotherapy students in the work place.

2.2 Method

Construction of a new measurement instrument requires a guiding framework that fosters validity and reliability through each stage of instrument development. A comprehensive framework widely used in the development and evaluation of standardised tests/instruments in psychology and education is provided in the Standards for Educational and Psychological Testing (American Educational Research Association, 1999).

The development of the instrument reported in this thesis was guided by the Standards (1999) and incorporated the 'four building blocks' approach proposed by Wilson (2005). This four step process starts by defining and visually representing the construct to be measured (Step 1) and then moves to developing items that assess the construct (Step 2). These items then generate responses that are scored, designated by Wilson (2005) as the outcomes space (Step 3). In Step 4 the measurement (statistical) model is applied to analyse the scored responses and to examine how well the construct appears to have been measured. The sequence of the four building blocks is a cycle of action and reflection (Figure 2.1) which is repeated several times during pilot and field testing of the instrument (Wilson, 2005).

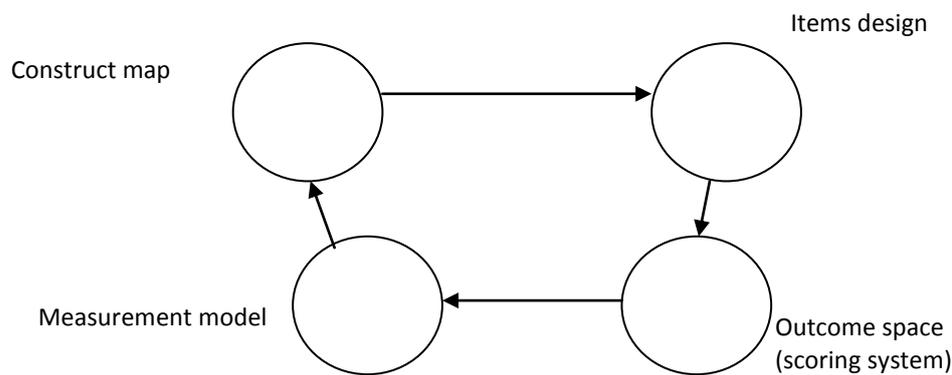


Figure 2.1: The instrument development cycle through the four building blocks (adapted from Wilson, 2005 p. 19)

2.2.1 Construct Mapping (Step 1)

Construction of an instrument commences with some preliminary conceptual decisions, that is, by defining the construct or trait that is to be measured (Streiner & Norman, 2003; Wilson, 2005). A construct map is a visual tool for clarifying the underlying construct to be measured by the instrument. The main idea underlying the use of a construct map during the initial stages of instrument construction is for the developer to focus on the essential feature of exactly what is to be measured, that is, to differentiate what is to be measured from other closely related but different constructs (Wilson, 2005).

A construct can be most readily expressed as a construct map where the construct has a single underlying continuum – “implying that for the intended use of the instrument, the measurer wants to array the respondents from high to low or left to right in some context” (Wilson 2005 p26). An example of a construct map is shown in Figure 2.2. The central arrow represents the underlying continuum of the construct to be measured (in this example, physical functioning) while the left hand side maps the respondents in order from more to less and the right hand side maps the item responses ordering them again from more to less. The construct map attempts to make the extremes of the continuum more concrete by describing them in detail which enables the intermediate levels to also be described.

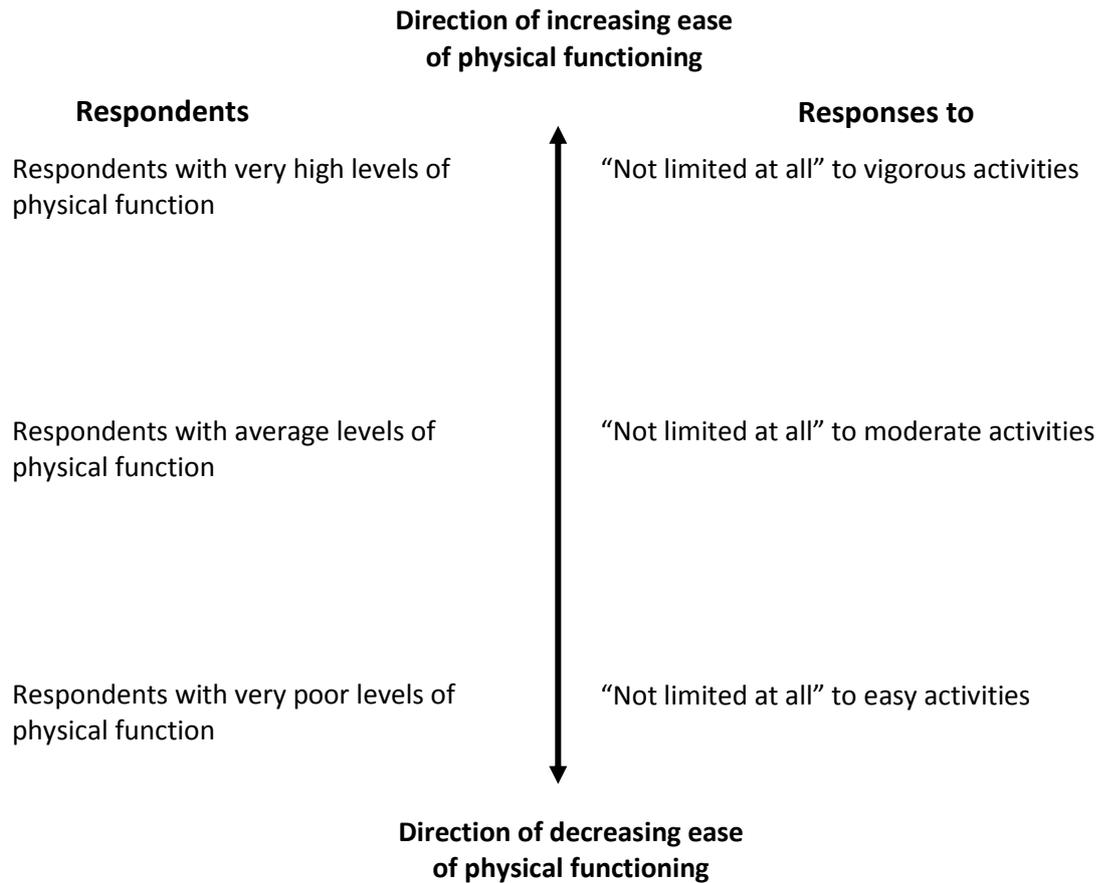


Figure 2.2: Construct map for physical functioning (adapted from Wilson, 2005 p. 31)

2.2.2 Items design (Step 2)

Because professional competency is observed through a variety of manifestations, rather than directly, it is considered to be a latent (hidden) trait or construct. The next step in instrument development is to generate a pool of items that are representations of the construct of professional competence.

2.2.2.1 Assembling an item pool

A comprehensive pool of items can be assembled by collating items from multiple sources including pre existing instruments, other relevant documents (e.g., professional standards), theory and relevant research, and gathering the views of experts in the field and a

comprehensive cross section of stakeholders (American Educational Research Association, 1999; Streiner & Norman, 2003; Wilson, 2005).

A large pool of assembled items enables item reduction based on a pre-determined set of criteria applied by independent reviewers. The quality of the final item set can be developed and refined through iterative independent review procedures, for example, focus groups, key informant interviews and pilot trials of the instrument on the target population. In addition, mapping of the items against the relevant professional standards enables comprehensive and detailed blueprinting of items against the construct (professional competence) being measured.

2.2.3 Outcome space (scoring system) (Step 3)

Wilson (2005) defined an outcome space as “any set of qualitatively described categories for recording and/or judging how respondents have responded to items” (p. 63).

Development of a scoring system involves decisions regarding how to categorize observations of an assessor and then score them so that the obtained value provides a valid measure of the construct (in this research, professional competence). The scoring procedure chosen must be consistent with the purpose and context of the test, well defined, finite, exhaustive, ordered, and facilitate meaningful score interpretation (American Educational Research Association, 1999; G. N. Masters & Wilson, 1997). The right side of the construct map in Figure 2.2 provides the beginning stages for development of a proposed scoring system.

2.2.3.1 Purpose and context of the assessment

Assessment of habitual performance in the clinical environment is essential for making judgments about clinical competence and professional behaviours and importantly, for guiding students towards expected standards of practice performance (Govaerts, van der Vleuten, & Schuwirth, 2002). As outlined in section 2.1.1, longitudinal assessment of professional competencies in the workplace is a component of all physiotherapy education in Australia. Therefore, in this research, the design of the assessment instrument was guided by the knowledge that it would be used in the workplace by clinical educators in their dual

roles of both facilitating and assessing student learning. Educators would be expected to judge and rate the student's ability to perform expected professional competencies at a defined standard.

2.2.3.2 Score interpretations – criterion or norm referenced assessment

Since the purpose of work-based assessment is to judge the student's ability to demonstrate professional competencies in relation to a specific standard rather than in relation to the performance of other students, the assessment is defined as criterion referenced. The standard is a fixed reference point and provides a decision rule for when an acceptable level of performance has been achieved, and will differ with the objectives of the clinical experience. Sometimes objectives are 'set' by accreditation bodies but invariably there are details in the assessment process that must be resolved through a consensus process with academic institutions and members of the profession involved in the education and assessment of students (van der Vleuten, 1996b). In Australia, the accrediting body does not provide decision rules regarding when students have achieved an acceptable standard of practice across the desirable competencies. A challenge that was identified early in this research was how to define the standard against which students should be assessed, given that the spectrum of experience to be assessed would range from the student making their first foray into the world of clinical practice to a student about to gain entry level qualification. It was anticipated that recognition of an acceptable standard of practice might require substantial negotiation with the practicing community. The problems associated with gaining profession wide agreement on ratings of competence would be amplified if multiple standards (for novices to expert students) were attempted. Conversely if a single standard was targeted for consensus (e.g. entry level competence), this would be challenging for the student and educator in early clinical experiences.

2.2.3.3 Data Types – Scaling

Stevens (1946, 1951) proposed that measurements can be classified into four different types of scales. These are:

- Nominal: numbers are used as labels, they express no mathematical properties, for example, gender

- Ordinal: numbers indicate the relative position of items, but not the magnitude of difference, for example, a class rank. Someone ranked at four has a higher GPA than someone ranked as five, but the difference between a four and a five is not quantified.
- Interval: numbers indicate rank, separated by equal intervals on a continuous scale, but zero does not indicate complete absence of any of the target variable (e.g. temperature in centigrade)
- Ratio: numbers indicate rank, separated by equal intervals on a continuous scale and there is a fixed and meaningful zero point on the scale (e.g. mass, length). Ratios calculated using these data are interpretable (e.g. twice as long, twice as much money).

The level of performance-based competence exhibited by a student in a complex workplace environment is likely to be continuous rather than categorical (Prescott-Clements, et al., 2008; Streiner & Norman, 2003). Categorical judgments are usually dichotomous (can do it, can't do it) and provide nominal measurement data. An example of a categorical judgment might be pass or fail. Continuous variables are of interest when measuring changing performance and provide ordinal measurement data. Categorising/dichotomising a continuous variable reduces the efficiency of an instrument as it prohibits the tracking of skill development for a fail student until a critical cut point is reached and the student is deemed to pass. Data on an ordinal scale provides students with feedback about relative improvement in skills across time, and allows ranking of students from strongest to weakest to identify students in need of tailored support and those who might be eligible for awards, scholarships, advanced training or specialisation (Hunter & Schmidt, 1990; Streiner & Norman, 2003). It was anticipated that a scoring system for assessing competence of physiotherapy students would allow ranking and feedback to students regarding the progression of their skills and that the data would be ordinal rather than interval in nature.

Physical format of the scale for a continuous variable

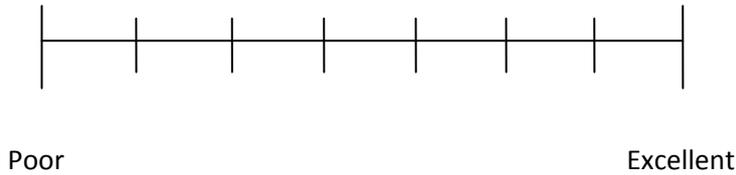
When considering rating the performance of a student within the clinical environment the most commonly used approach in physiotherapy is the direct estimation technique (Spector,

1992; Streiner & Norman, 2003). Direct estimation techniques are designed to elicit from the assessor an estimate of the magnitude of the construct being measured. When considering assessment of performance, there are three useful formats for summated rating scales designed to achieve direct estimation, the visual analogue scale (VAS), the adjectival scale and the Likert scale (Kline, 1986; Spector, 1992; Streiner & Norman, 2003). These formats may also be used in combination to describe and judge competency, for example, the Behavioural Observation Scale (BOS). A VAS is usually a horizontal (but may be vertical) line, 100 mm in length, anchored by word descriptors at each end. A range of rating scale options are illustrated in Figure 2.3.

Visual analogue scale



Discrete visual analogue scale

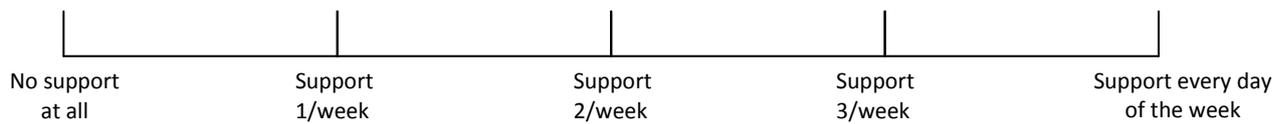


Discrete visual analogue scale (Wong & Baker, 1988)



Adjectival scale

What level of support do you need to be able to continue living at home?



Likert scale

Please indicate how much you agree or disagree with each of these statements:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Student acts on feedback	1	2	3	4	5
Student is punctual	1	2	3	4	5

Figure 2.3: Examples of rating scales: VAS, DVAS, Adjectival and Likert.

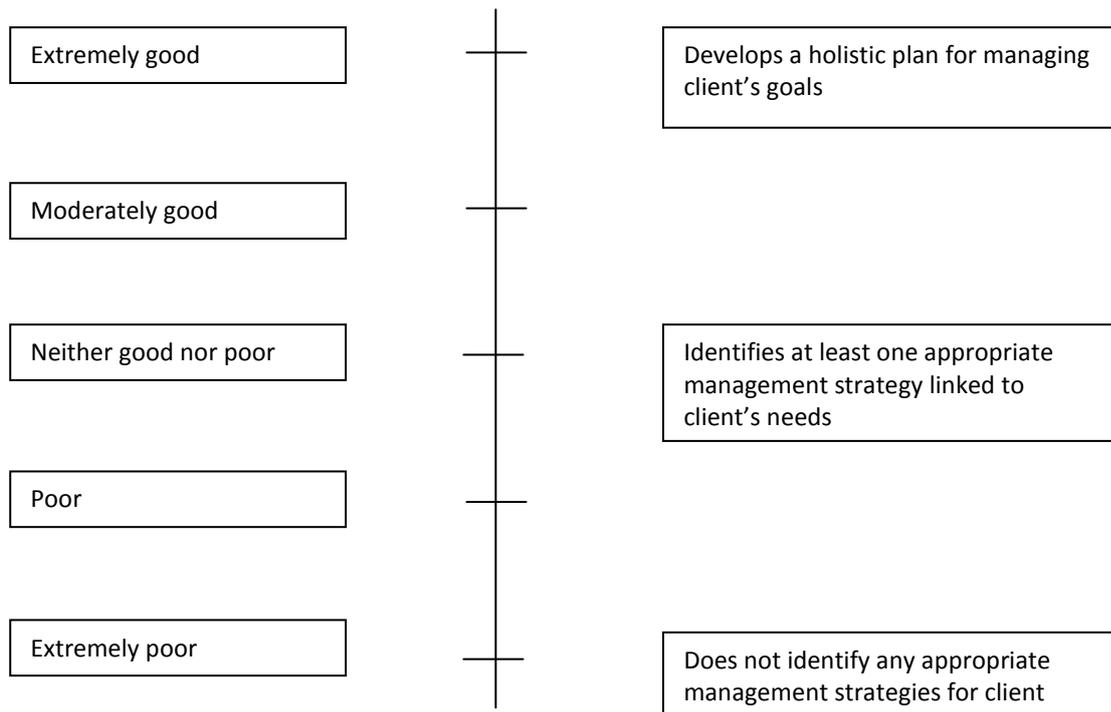
The assessor places a mark anywhere along the line that matches their judgment of the ratee's performance on the continuum. A VAS score can be determined by measuring in millimetres from the left hand end of the line to the point that the assessor marks. If a VAS has pre-specified points marked on the line between the two anchor points (Figure 2.3), it is termed a discrete visual analogue scale (DVAS).

Adjectival rating scales are very similar to the VAS but have descriptors placed at intervals along the line to guide the assessor's rating (Landy & Farr, 1980) (Figure 2.3). More commonly a set of check boxes rather than adjectives on a line are used in adjectival scales. The VAS, DVAS and adjectival scale are considered unipolar scales which prompt the assessor to think of the presence or absence of an attribute (scale values range from 0 – 100% of the attribute). Unipolar adjectival scales commonly use a 5 point numeric scale choosing a label for each point, for example, not at all satisfied, slightly satisfied, moderately satisfied, very satisfied, extremely satisfied.

The Likert scale is a bipolar scaling method, measuring either positive or negative response to a statement. In Likert scales the end points are opposites, such as "completely dissatisfied" and "completely satisfied". There are usually an odd number of response options with a neutral option in the centre. Response levels are labelled and anchored with consecutive integers as illustrated in Figure 2.3 (Streiner & Norman, 2003).

There are numerous combinations of these three rating scale formats. Three examples are illustrated in Figure 2.4 and include behaviourally anchored scales (BAS), behavioural observation scales (BOS), and global rating scales (GRS).

Behaviourally anchored scale



Behavioural observation scale

1. Establishes short term goals for management of client

Rarely 1 2 3 4 5 Always

2. Seeks assistance where appropriate

Rarely 1 2 3 4 5 Always

Global rating scale

The overall performance of the student in this clinical placement was:

Not adequate Adequate Good Excellent

Figure 2.4: Examples of rating scales: BAS, BOS and GRS

The other option is to rate the subject according to the degree to which they possess a trait in relation to particular competencies e.g. Trait/Global, VAS or adjectival rating scales. All of these approaches generally involve specifying the behaviours to be observed in some detail.

Due to the number of possible scaling options the VAS offers apparent precision while at the same time appearing easy to complete. It has been demonstrated that there is only an illusion of precision, as little information is lost if scales with fewer grading options are used. This is because most people appear to collapse the 100mm line into a 5-7 point scale (Jensen, Turner, & Romano, 1994; Straube & Campbell, 2003). Wording of the end points can cause different interpretations of the scale between respondents, while older people and people with low levels of literacy have been shown to have difficulty completing a VAS (Bosi Ferraz, et al., 1990; Krosnick & Presser, 2009; Seymour, Simpson, Charlton, & Phillips, 1985). There is potential for a halo effect if items are presented in a single column on one page, where they may not be rated individually but rather on the basis of a global impression (Streiner & Norman, 2003). When considering assessment of professional competence, a unipolar scale is often the scale of choice as its structure more closely represents the underlying continuum of performance from very poor (incompetent) through to very high levels of competence, with individual students demonstrating more or less of the variable (Wilson, 2005).

Once it is determined that a performance assessment is required, the physical format of the rating scale may not be as critical as other factors. Kingstrom and Bass (1981) reviewed research comparing scale formats including VAS, Likert styles, or mixed formats. They concluded that there was little or no difference between the psychometric characteristics of measurements obtained using the different scale formats. Similarly Landy and Farr (1980) reviewed performance assessment and concluded that scale format explains only 4 to 8% of score variance. More important factors appear to be the context in which the scale is used, its specific purpose and the comfort of stakeholders with the final decision.

2.2.3.5 Further decisions in construction of a continuous scale

As well as the physical format chosen for the scale, there are additional issues that must be addressed in the design of a scoring system to maximize precision, minimize bias and target adequate validity.

Number of points on the scale: How many is appropriate?

Miller (1956) suggested that seven levels, plus or minus two, are the finest degrees of perceptual discrimination humans can typically make in any situation. A greater number of response options may be redundant and misleading. As previously proposed, even with a VAS people tend to reduce a 100 point scale down to five to seven categories (Cicchetti, Shoinralter, & Tyrer, 1985; Fay & Latham, 1982; Munshi, 1990; Preston & Coleman, 2000; Straube & Campbell, 2003; Streiner & Norman, 2003). In all scales inter-rater reliability may be best with seven to ten categories and not improved by increasing the number of categories beyond ten. To explore the relationship between scale length and reliability Krosnick (1991) conducted a meta-analysis of 706 studies. They found that five- or seven-point scales produced the most reliable results and proposed that measurement error may be introduced by 'noise' created by more categories. In unipolar scales some studies found no relationship between the number of scale points and reliability (Matell & Jacoby, 1971). Others have found that reliability is greater for 4 point compared to 2 point scales (J. R. Masters, 1974; Watson, 1988), 5 point compared to 7 or 11 point scales (McKelvie, 1978) and 5-7 point compared to 3 or 9 point scales (Matell & Jacoby, 1971).

In the case of unipolar scales it appears that too few levels on the scale and information is lost; too many and the assessors cannot discriminate between scale levels. Krosnick and Farbrigar (1997) stated that people can readily conceive of zero, a slight amount, a moderate amount and a great deal along a unipolar continuum. Hence they recommended that an optimal design would be to have 4 – 7 response options. These authors also argued that raters tend not to use the end points of a scale and this needs to be taken into consideration when developing and evaluating a rating scale. It is also important to determine how a rating scale is being interpreted by raters, make appropriate modifications to the scale and evaluate the function of the scale prior to using the data to make decisions.

Should there be an even or odd number of scoring categories?

In unipolar scales having an even or odd number of scoring categories is irrelevant. In bipolar scales (like strongly agree – strongly disagree), scales with an even number of categories force the respondent to make a choice, whereas, scales with odd numbered categories can offer the option of having no opinion by selecting a neutral midpoint between two extreme scores (Krosnick & Farbrigar, 1997; Krosnick & Presser, 2009; Streiner & Norman, 2003). In these scales use of a mid-point can be justified if it is believed that the respondents genuinely may have a neutral position. Alternatively though, respondents may opt for the neutral mid-point because they are unsure of how to rate and it is an easy choice requiring little cognitive defence.

Unbalanced scales, where neutral is not at the mid-point may produce bias allowing a scale to produce relatively more positive or negative data (Spector, 1992). However they can be useful if there is likely to be an overwhelming response in a specific direction.

Research has shown that people do not generally choose mid points if the assessment outcome has a high level of importance to them (Farbrigar, Krosnick, & MacDougall, 2005; Krosnick & Farbrigar, 1997; Schuman & Presser, 1981). No consistent pattern has been identified between the presence of a mid point and reliability of the scale (Malhotra, Krosnick, & Thomas, 2007).

Overall an odd or even number of categories appears to be of little consequence and the number of scoring categories is best decided in relation to the population using the scale and their needs/preferences (Krosnick & Farbrigar, 1997).

Labelling of scales and the influence of wording on scale interpretation

Another important decision in scale construction is whether to label some or all points with words or numbers or a combination. Due to the potential for ambiguity with verbal labels, numerals may offer more precision. However, since people rarely express their thoughts in numerals, choosing a verbal label may be a more natural mental activity than selecting a number within a range (Krosnick & Berent, 1993; Spector, 1992). When number options are presented, their interpretation needs to be explained with written labels. Placing the verbal description on the scale circumvents the need for this reference text. While respondents preferred rating scales with more verbal labels (Dickinson & Zellinger, 1980), if only some

boxes or points on the scale are labelled there is a tendency for these points to be chosen in preference to unlabelled points (Streiner and Norman 2003). Christian et al.(2009) found that using fully labelled scales provided greater inter-rater reliability, however this occurred more so in respondents with low to moderate education.

Schwarz et al.(1991) conducted a series of experiments investigating the influence of numerals on a person's interpretation of score labels. In a self-administered questionnaire, two groups of respondents were asked to report, along 11-point rating scales, how successful they have been in life and how happy a childhood they had. One group used a scale ranging from 0 to 10 and the second group from -5 to +5. The endpoints were labelled "not at all successful" and "very successful" or "unhappy" and "very happy". Coding both scales from 0 to 10, respondents reported higher success in life for themselves (M = 7.38) along the -5 to +5 scale than along the 0-10 scale (M = 5.96), resulting in a pronounced main effect for numeric values ($F[1,93] = 16.21, p < .001$). Similarly, respondents reported higher childhood happiness along the -5 to +5 scale than along the 0-10 scale. These results indicate that the numerals on the rating scale may have moderated how the participants interpreted the scale labels.

Similarly Lam and Kolic (2008) found that there needs to be congruence between the content of items and scale labels or the reliability and validity of performance ratings can be affected. They recommended that scale anchors should reflect the wording of the items. Rating scales that exhibit semantic incompatibility tended to decrease rater reliability in assessment of performance. Willis (2005) and Lam and Kolic (2008) also recommended that qualitative data is collected through interviews to understand how and why people respond to items and scales as they do.

Performance indicators: Are behavioural indicators of performance required?

As discussed in section 2.2.1, longitudinal assessment of professional competencies in the workplace is practiced in all physiotherapy education in Australia with clinical educators having dual roles of both facilitating and assessing learning. Clinical educators are expected to judge and rate the student's ability to perform expected professional competencies. The Standards (1999) recommend that assessors are provided with examples of behaviours on

which to base these judgements. Further, performance indicators that describe expected behaviours need to be unambiguous, observable, measurable, guide improvement in performance, and be transparent to both students and assessors (J. Cox, 1996; Krosnick & Presser, 2009).

Loomis (1985a) reviewed the medical education literature and concluded that to improve rater reliability, competencies and associated performance standards need to be well defined in terms of observable behaviours or standards that describe the levels of mastery of the competency. Providing examples (performance indicators) of expected behaviours for each item on an instrument enables all parties to be clear about the elements in performance that require attention and convert these to achievable goals (formative assessment). Performance indicators also provide benchmarks that aid decisions regarding whether the student's performance has reached the minimum acceptable standard to progress in their program of study (summative assessment). Thus performance indicators and the rating scale form an integrated scoring system when assessing performance.

Is the scoring procedure analytic or holistic?

Two types of scoring procedures can be used: analytic and holistic. Analytic is the scoring of individual items, that is, each essential element of the performance is assessed independently; separate item scores are obtained as well as an overall total score. Holistic scoring consists of a global rating of performance (American Educational Research Association, 1999). The analytic approach provides specific information on each aspect of performance whereas holistic scoring procedures may be preferable when an overall judgement of performance is required. This is usually when the skills being assessed are complex and highly interrelated. Research in medicine suggests a combined or hybrid approach utilising both specific items and global rating scales may be the most suitable method of evaluating undergraduate clinical performance to draw on the strengths and address the limitations that are observed when the rating scales are used in isolation (McIlroy, Hodges, McNaughton, & Regehr, 2002; P. J. Morgan, Cleave-Hogg, & Guest, 2001). In objective structured clinical examinations (OSCE), itemised checklists and global rating scales show similar levels of inter-rater reliability (D. Newble, 2004). The key issue is the

context in which the scale is being used, its specific purpose and the comfort of stakeholders with the final decision.

2.2.4 Measurement model (Step 4)

The next step in development of an instrument is to evaluate the way that scores provide evidence of the underlying construct. This is done through the fourth building block, the measurement model, that is, the psychometric or statistical model chosen to evaluate scores provided by the instrument. The object of the measurement model is to analyse the scored responses, to examine how well the construct appears to have been measured and guide the interpretation of the results and their practical applications (Wilson, 2005).

2.2.4.1 Theories underpinning instrument development and evaluation: Classical Test Theory (CTT)

Classical Test Theory (CTT) has traditionally been the main paradigm for the design, analysis and scoring of tests and questionnaires. CTT is based on three concepts, test score (X) (often called the observed score), true score (T) and error score (E) where $X = T + E$. For each examinee there are two unknown variables (T and E); the equation cannot be solved unless some assumptions are made (Hambleton & Jones, 1993; Novick, 1966):

1. error scores are uncorrelated with each other and with the true scores
2. the average error score in the population is zero
3. error scores on parallel tests are uncorrelated
4. observed, true and error scores are linearly related

However, because true scores and error scores cannot be determined, the appropriateness of the assumptions cannot be verified and it can only be postulated that they are met.

There are several significant problems with the methods used by traditional psychometric approaches in handling data, and constructing and evaluating scales. The most relevant of these are outlined below.

1. Ordered counts are not interval measures

Most rating scales used to assess work-based performance use the method of summated ratings. Here, rating scale options are allocated sequentially ordered integers, and item

scores are summed to give a total score. While this approach is common, there is little evidence to support the proposition that ordinal level total scores approximate interval-level measurements (Cliff & Keats, 2003; Michell, 2008; Streiner & Norman, 2003).

2. Item and scale statistics are sample dependent

This limitation of CTT means that the item and scale statistics apply only to the specific group of subjects who took the test. This means that if the scale is to be used on a different population then it is necessary to re-establish the psychometric properties of the instrument in the new population. Similarly if any of the items were altered or removed, reestablishment of the psychometric properties is necessary.

3. Missing data

In CTT missing item scores are replaced with the person-specific mean score (the mean score of the items answered by that individual). This is appropriate if items all have the same level of difficulty. Realistically estimating how a person will score on any given item is not reasonable given that each item has a different propensity to measure the underlying attribute (construct) and is designed to assess a range of desirable attributes in a competent practitioner.

4. Standard error of measurement (SEM)

Using CTT the error associated with a person's score (SEM) is commonly treated as a constant value, that is, the error of measurement is the same at the ends of the scale, where scores are typically least precise, as in the middle of the scale where scores are most precise (Stratford, et al., 1996). The SEM for item and test scores can be large (and precision low) when estimates are based on relatively small samples with a wide spectrum of student ability.

5. Scaling items

When a set of items make up a measurement instrument, it is desirable to develop a common continuum on which student ability and items difficulty are located. Thus, when a set of items is used to measure some trait of a person there is an interaction between person ability and item difficulty. Traditional psychometric methods however do not scale items, that is, they do not enable assessment of item difficulty and do not enable them to

be located on a measurement continuum. All items are considered equally difficult and typically contribute equally to the total score.

2.2.4.2 Theories underpinning instrument development and evaluation: Item Response Theory (IRT) and Rasch Measurement Model (RMM)

Alternative approaches for the design, analysis and scoring of an instrument are Item Response Theory (IRT) and the Rasch Measurement Model (RMM). These two approaches investigate a respondent's true measurement on the construct being measured by the scale (i.e. their location on an interval-level continuum) by converting total scores, which are ordinal in nature, to data with interval level properties. Additionally it is a person's true interval-level location that predicts their ability and or score on each item.

There are a number of important differences between IRT and RMM. In the paradigm of IRT models, the emphasis is on finding a model that best characterizes the given data while the RMM places greatest importance on the inherent properties of the mathematical model (Bond & Fox, 2007). The emphasis is in identifying and studying anomalies in the data disclosed by the Rasch model. When the data do not fit the Rasch model another model is not chosen. Instead the data is examined to determine why the instrument is not performing as anticipated and modified to achieve an ideal scale. The ability to estimate the item and person locations independently of each other is a feature only of Rasch models (Andrich, 1988). The RMM enables estimation of item difficulty and student ability. The model predicts that a student with known ability x should be able to successfully complete items with difficulty levels lower than x and struggle with successful completion of items with difficulty level greater than x . This approach enables test developers to create a scale where higher scores indicate a greater ability with respect to the underlying construct; educators can identify challenging items and appropriate educational support can be developed to help students achieve more challenging targets; and the scale can be evaluated and revised so that predicted responses (based on known student ability and item difficulty) are evident in student outcomes.

An additional critical advantage of Rasch measurement is that it enables the abstraction of equal units of measurement from the raw data of observations i.e. scores on the items of an assessment tool. These can be calibrated and then used with confidence to measure human attributes such as competence in physiotherapy practice (Bond & Fox, 2007; G. N. Masters & Keeves, 1999). This conversion facilitates appropriate interpretation of differences between individuals and allows for tallying of scores to provide a meaningful total test score. In addition, Rasch analysis enables testing of the internal construct validity of a scale for unidimensionality (considered an essential quality of a scale scored by adding results of items) and identification of gaps in the targeting of items to students' abilities. Rasch analysis also enables assessment of differential item functioning (DIF). DIF testing investigates whether bias in scores occurs for specific subgroups. This might occur if test results varied with clinical area of practice or if males and females were rated differently using the scale (Tennant & Conaghan, 2007). Dichotomous, Rasch (1960), and polytomous, Andrich (1978), versions of the Rasch model are available. In summary, Pallant and Tennant (2007) suggest Rasch analysis enables formal assessment of the measurement properties of an instrument through investigation of the following: overall model fit, overall person fit and item fit, individual item fit, thresholds, targeting, person separation index (PSI), differential item functioning (DIF) and local independence (dimensionality).

It is argued that the Rasch Measurement Model (RMM) is the standard for psychometric evaluations of outcome scales, and should be used during the development phase or when reviewing the psychometric properties of existing instruments (J. Hobart & Cano, 2009; J. C. Hobart, Cano, Zajicek, & Thompson, 2007; Tennant & Conaghan, 2007). The systematic review, (presented in Chapter One), revealed that only three of the eight instruments located had been investigated (incompletely) using IRT or RMM. The CPI and PT MACS were investigated by Logemann (2006) using a 2-parameter graded response IRT model and the SCCS using Rasch analysis by Rheault and Coulson (1991).

2.3 Action Research: Synthesis of instrument development

The sequence of instrument development through the four building blocks is a cycle of action and reflection which is repeated several times during pilot and field testing of the

instrument, enabling continuous evaluation and modification (Wilson, 2005). This approach facilitates the development of an instrument with comprehensive evidence of validity. One of the key steps is engagement with and participation of a broad cross-section of stakeholders at each stage of instrument development. Participation of stakeholders promotes buy-in, ownership and uptake of the instrument (Cross & Hicks, 1997). The approach illustrated in Figure 2.1 can be extended to include the iterative cycles of instrument development, pilot and field testing and continuous refinement of the instrument based on evaluation throughout the different phases as shown in Figure 2.5.

This cyclical research paradigm of planning, acting, monitoring and evaluating is best described as participatory action research and has been used effectively by occupational therapists and speech pathologists to develop a process to assess fieldwork performance of students (Allison & Turpin, 2004; McAllister, 2005). To achieve adequate rigour it utilises a reflective spiral. Each turn of the spiral integrates theory and practice, understanding and action, and informs the next turn.

When developing an instrument for assessment of work-based performance, the research situation demands stakeholder participation, collaboration and responsiveness. Checkland and Holwell (1998) suggest action research best incorporates these demands. It is, above all, a method for yielding simultaneous action (change) and research outcomes. In the context of assessment of competence to practice physiotherapy, a desirable outcome would be a high quality assessment instrument that was willingly adopted by all programs in Australia and New Zealand. The results of this type of research are practical, relevant, and improve professional practice through continual learning and progressive problem solving.

2.4 Instrument development: quality assurance processes

Within this proposed iterative research framework (Figure 2.5), planning the quality assurance processes necessary at each step enables stakeholders to be confident that measurement validity has been addressed appropriately and the instrument is fit for its intended purpose. There are a number of ways to strengthen confidence in the final instrument. These quality assurance processes are summarised in Table 2.1.

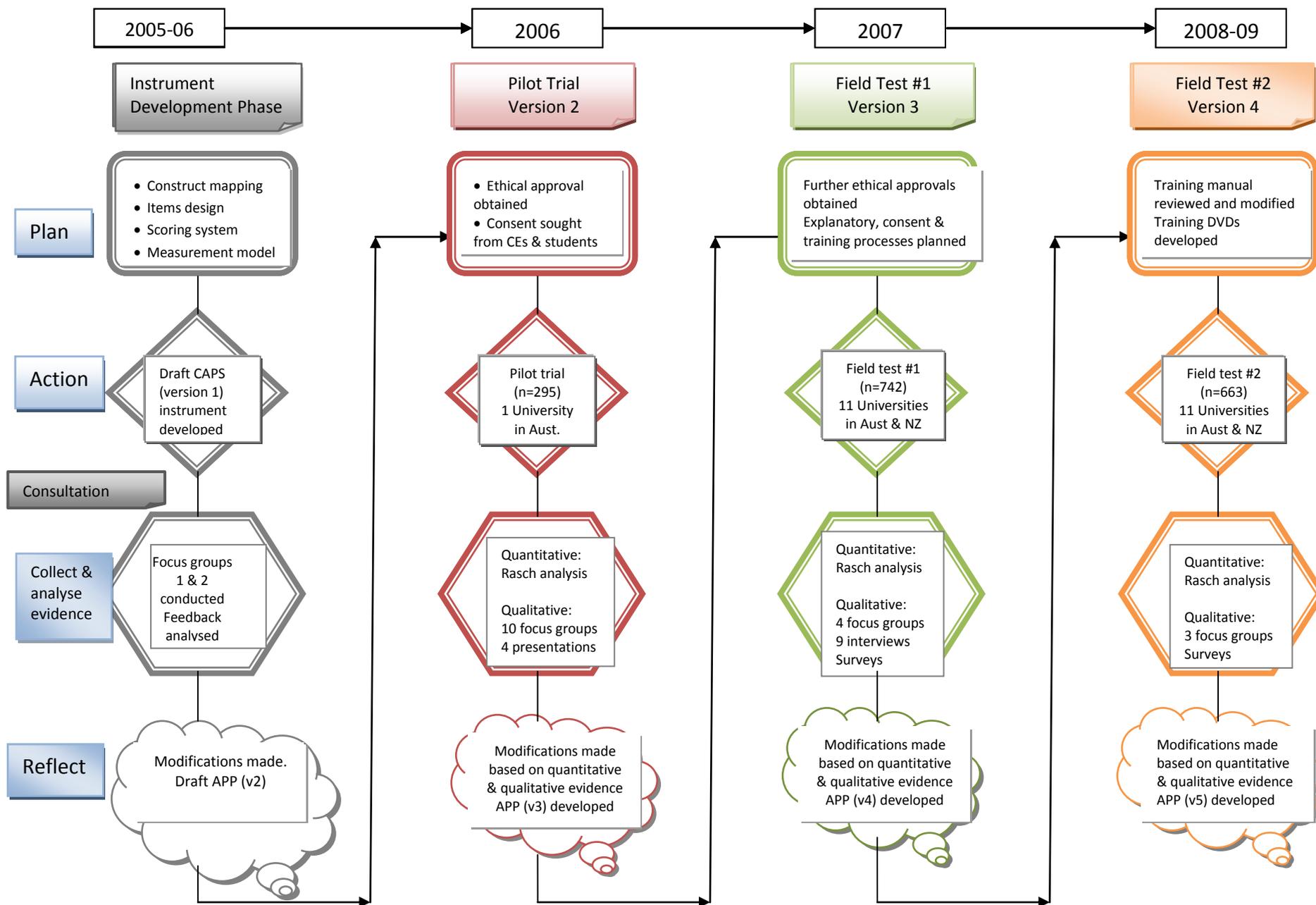


Figure 2.5: Iterative research framework

Table 2.1: Summary work-based assessment instrument development: quality assurance processes

Criteria	Quality Assurance Processes
Construct map	<ul style="list-style-type: none"> • Discussion with purposively sampled experts including research team • Draw map • Present to stakeholders for feedback and modification
Items design	<ul style="list-style-type: none"> • Collate items from all existing instruments, relevant documents, theory and research • Assemble item larger item pool than required • Engage with experts in the field & all relevant stakeholders • Ensure comprehensive cross section of users is engaged in process of development • Standardised criteria for item reduction defined: technical as well as content • Independent item reduction (of initial item pool) • Ensure quality through independent review processes of items eg focus groups etc and pilot trial/field tests • Evaluation of the extent to which the items match the definition of the construct and the purpose of the test. E.g., Map against relevant professional standards • Specify test specifications
Scoring system	<ul style="list-style-type: none"> • Review purpose and context of the assessment • Review all available scoring systems in use in current instruments • Decision on norm vs criterion referencing • Decision on level of measurement (categorical or continuous) • Decide on appropriate scale format • Ensure quality through independent review processes of scoring system eg focus groups etc and pilot trial/field tests • Development of appropriate scoring guides (eg training manual). • Examples of behaviours representative of the level of performance being measured (performance indicators)
Measurement model: (ensure model chosen provides data on these aspects of instrument development and evaluation)	<ul style="list-style-type: none"> • Evidence on instrument internal structure • Differential item functioning • Discriminatory ability of rating scale • Reliability • Identification of gaps in item content, redundant items • Dimensionality of instrument • Targeting: floor or ceiling effects • Sample and item independence • Results for missing item data handled scientifically
Pilot trial and Field tests	<ul style="list-style-type: none"> • Participatory action research approach • Representative sample of target population • Representative sample of end users • Appropriate measurement model to ensure adequate quantitative data for analysis • Independent qualitative review processes of pilot trial/field tests (e.g., focus groups, questionnaires, interviews) • Triangulation of qualitative data through: replication of results, Interviews, focus groups, workshops, emails, tele/video conference

2.5 Instrument development: Summary

The systematic review presented in Chapter One found a range of research of variable design and methodological quality examining eight instruments developed to assess performance of physiotherapy students within the clinical environment. The review failed to identify convincing evidence sufficient to support adopting an existing instrument for use by physiotherapy programs in Australia and New Zealand. The results of the systematic review became the driver for the development and evaluation of a new assessment instrument. Construction of a new instrument requires a guiding framework to ensure all aspects of validity are adequately addressed. In this chapter the framework provided by the American Educational Research Association (1999) and Wilson (2005) is outlined, including mapping of the construct to be measured, design of the items and scoring system and choice of a measurement model to guide development, refinement and evaluation of the new instrument.

One of the key components of successful instrument development is engagement with and participation of a broad cross section of stakeholders at each stage of development. The cyclical research paradigm of participatory action research addresses the demands of participation, collaboration and responsiveness. Forward planning of quality assurance processes necessary at each stage of instrument development enables stakeholders to be confident that validity has been addressed appropriately and the instrument is fit for its intended purpose. This Chapter describes and defends a proposed plan for instrument development, and is followed by Chapter Three, which details the steps undertaken in development of a new instrument to assess professional competence of physiotherapy students, and presents the first version of this instrument.

3. Chapter Three: Development of the Assessment of Physiotherapy Practice (APP) instrument

3.1 Introduction

Chapter Three applies the theoretical framework described in Chapter Two to the development of an instrument suitable for assessing competence of physiotherapy students.

3.2 Methods (Part 1): Instrument development phase

Instrument development was the first stage of the iterative research cycle illustrated in Chapter Two, Figure 2.5.

3.2.1 Project Team and funding

A small group of key personnel was assembled to monitor and inform the process of instrument development (Streiner & Norman, 2003; Wilson, 2005). The research team consisted of three academics with a broad range of experience pertinent to clinical education and assessment of professional competence within physiotherapy (JK, MDAL and MDAV) (see Appendix 3.1). In addition, the 16 members of the Australian and New Zealand Universities Physiotherapy Clinical Education Managers group acted as a broader reference group throughout the research (see Appendix 3.1). Funding for this project was provided in part by a grant from the Australian Learning and Teaching Council, an initiative of the Australian Government Department of Education, Employment and Workplace Relations.

3.2.2 Aims of the Research

The primary aims of the research were to:

- develop an assessment instrument to evaluate the clinical performance (professional competence) of physiotherapy students in the workplace
- investigate and advance the psychometric properties of the instrument
- investigate the feasibility/utility of the instrument when applied to measure competency in authentic practice settings.

3.2.3 Construct Mapping

The research group agreed that the underlying construct to be measured by the new instrument was 'professional competence', and drafted a map for consideration. Professional competence was considered a continuum of performance from very poor (incompetent) through to very high levels of competence. In entry-level physiotherapy education, student performance is assessed by a clinical educator (CE). The construct map included a tentative proposal for how an educator might classify student performance.

3.2.4 Item design

3.2.4.1 Assembling an item pool

To generate a pool of items that could be considered for inclusion in a new instrument, a comprehensive item pool was assembled by drawing items from a broad range of relevant sources including all instruments in use in Australian and New Zealand physiotherapy programs, Australian Physiotherapy competency standards (ACOPRA, 2002), Australian Standards for Physiotherapy (Australian Physiotherapy Council, 2006), National Patient Safety Framework (Australian Council for Safety and Quality in Health Care, 2005), National Occupational Therapy competency assessment document (Allison & Turpin, 2004), National Speech Pathology competency based assessment tool, COMPASS™ (McAllister, 2005), and the Australian Council on Healthcare Standards (Australian Council on Healthcare Standards, 2002). Items were extracted and assembled in an excel spreadsheet. Two independent assessors (MDal and JK) identified duplicate items and retained the least ambiguous version. Disagreements were resolved by discussion. Where necessary a third reviewer's opinion was sought (MDav).

3.2.4.2 Item reduction

As a finite and relevant number of assessment items were required for practical assessment of clinical skills, a parsimonious item set was considered desirable. Item reduction was achieved by application of specific criteria (Krosnick & Presser, 2009; Streiner & Norman, 2003). Included items had to target one attribute (explicit learning outcome); describe an observable and measurable behaviour; be unambiguous, clear and defensible; be important to students, educators and/or key stakeholders; be described without jargon; be without

value-laden words, for example, the term trivial in “do you often visit the physiotherapist with trivial symptoms?”; be concise, as validity coefficients tend to fall as the number of letters in an item increases (Holden, Fekken, & Jackson, 1985); be free of negative wording e.g. ‘not’ or ‘never’. Item reduction was performed by two independent assessors, and where necessary a third reviewer’s opinion was sought. Disagreements were resolved by discussion.

3.2.4.3 Mapping of the Standards

Mapping of items against relevant professional standards is recommended as it enables comprehensive and detailed blueprinting of items against the construct (professional competence) being measured. After item reduction and before the instrument was presented to the focus groups, broad mapping of the items to the Australian Standards for Physiotherapy (2006) was conducted to ensure key areas of the standards were covered. Following feedback from focus groups and prior to the pilot trial, a detailed mapping of the item content to the standards was completed.

3.2.5 Outcome space (scoring system)

Development of a scoring system involved decisions regarding how to categorize and score observations of an assessor so that obtained values provide valid measures of professional competence (Wilson, 2005). The right side of the construct map (Figure 3.1) provided a proposal for a scoring system for consideration by stakeholders.

Based on the steps recommended in Chapter Two for best practice in development of a new instrument for assessment of professional competence, there were several decisions required when designing the rating scale. These decisions included:

1. Defining purpose and context of the assessment
2. Determining the nature of the scoring system: criterion versus norm referenced assessment
3. Deciding on the level of measurement data: nominal, ordinal, interval or ratio
4. Designing the physical format of the scale

A scoring system, and theoretical defence for its characteristics, was developed by the research group for consideration by stakeholders.

3.2.5.1 Development of performance indicators (PIs)

For each item in the initial item set, a list of performance indicators (PIs) were developed drawing on the same source documents that informed item development. The performance indicators were a non exhaustive list of behaviours intended to serve as a learning guide for students and provide educators with examples of unambiguous descriptions of behaviours that would indicate competence for each item. The criteria for the performance indicators were that they must describe expected behaviours, were unambiguous, observable, measureable, must guide improvement in performance, and be transparent to all stakeholders (students, assessors) (J. Cox, 1996; Streiner & Norman, 2003). An initial set of PIs for each item were developed by the research group for consideration by stakeholders.

3.2.6 Measurement model

The Rasch Measurement Model (RMM) was the method planned for examining the nature of item responses and how well a single construct appears to be measured. Rasch analysis also enables conversion of scores to interval data through logit transformations and informs interpretation and practical application of assessment results (Wilson 2005). It provides a sophisticated method for scale development and its advantages over classical test theory (presented in Chapter Two, (section 2.2.4)) have been convincingly argued (Andrich, 1988; Bond & Fox, 2007; Waugh, 2005). Results of Rasch analysis of pilot trial data are presented in Chapter Four.

3.3 Methods (Part 2): Consultation phase

An important factor in developing a valid instrument that is acceptable to end users is engagement with and participation of a broad cross-section of stakeholders at each stage of instrument development.

Qualitative methods, such as focus groups, one to one interviews and surveys, enable input from relevant stakeholders during instrument development. Krueger and Casey (2000) recommend selecting the qualitative method that best serves the particular purpose and desirable outcomes of the study. The research group considered nationwide stakeholder participation in the development of the instrument a matter of critical importance. A target of the project was to maximise the sense of ownership of the profession in the final product as it would be used by the profession to assess competence to practice. Focus groups, one to one interviews, surveys, workshops, email and teleconferencing were scheduled across the duration of the research to gather input of as many stakeholders as possible. This approach enabled triangulation and reinforcement of decisions based on qualitative data obtained from multiple sources.

Focus group data was also used to develop unambiguous items for inclusion in surveys completed by students and clinical educators (Barbour, 2005; McLeod, Meagher, Steinert, & Boudreau, 2000). Focus groups were structured to elicit information on potentially problematic areas of professional practice and explore barriers to uptake of the instrument and identify suboptimal assessment practices that could be addressed through educator training (D. L. Morgan, 1988). Focus group moderators allowed participants to share and discuss ideas (Barbour, 2005). They were intended not only to identify ways to refine the assessment instrument but also to “empower” participants by inviting them to play an active part of the process of analysis (Kitzinger, 1995)

3.3.1 Focus groups

Two focus groups were conducted to gather feedback on the draft items, the proposed rating scale and the performance indicators.

3.3.1.1 Participants

Purposive sampling was used to achieve the input of relevant and qualified participants. Purposive sampling amplified homogeneity of focus groups facilitating interaction between participants with similar backgrounds and experiences. Participants were screened to enable representativeness of the larger stakeholder population. Recruitment was designed

to optimise representation of all stakeholders by location (metropolitan, regional/rural and remote), clinical area of practice, years of experience as a clinical educator/supervisor or manager, organization (private, public, hospital based, community based and non-government). The desirable maximum number of attendees was twelve, but as the interest of the profession in the research grew, this maximum number was exceeded on occasion. Given the importance of stakeholder support and input, a decision was made not to exclude any suitable and interested participants. On the occasions when participant numbers exceeded twelve, an additional facilitator was used and the group divided into groups not larger than twelve for discussion.

3.3.1.2 Duration and site selection

Each focus group was scheduled for one and a half hours and arranged at a time and location to suit participants (Appendix 3.2).

3.3.1.3 Moderators

The principal moderator for each focus group was a research assistant with expertise in focus group methods. The research assistant was not a physiotherapist and had no vested interest in the outcomes of the research, reducing the likelihood of moderator bias. A second moderator took detailed notes, handled the logistics, such as refreshments, set up of the room, collection of consent, recording the focus group using digital recording devices and communicating with participants following the meetings.

3.3.1.4 Recruitment procedure

A potential participant pool was collated. This pool was based on knowledge by the research team of the set of individuals and organizations in Australia and New Zealand who arranged clinical education placements for physiotherapy students or who managed staff responsible for clinical education.

For each focus group initial contact with potential participants was made by email and/or phone and discussion included background information about the research, date/time and location of focus groups and an invitation to participate. Interested participants were sent

an information sheet, a consent form and the focus group agenda including an outline of topics for discussion (see Appendix 3.3). Follow-up phone calls were made two to three days after initial contact to confirm that potential participants had received the material for discussion and invitation to participate and to provide additional information requested by the potential participants.

An invitation to complete a survey, attend a teleconference or provide feedback via email at a later stage of the project was offered to all stakeholders who were unable to participate in a focus group.

3.3.1.5 Focus group protocol

If participants had not signed a consent form prior to attending the focus group, this was finalised prior to the group commencing. The focus group structure was outlined to participants, including the questions to be covered, that the focus group discussion would be recorded, transcribed and de-identified, the data was confidential and that participants were free to leave at any time.

3.3.1.6 Questions

Each group interview commenced with an introduction by the moderator and assistant moderator. The moderator reminded the group that anything said during the group interview should remain confidential, that the session would be audio taped and that the transcription would be de-identified. The moderator also informed the group that she was not a healthcare professional but someone with extensive experience in conducting group interviews. The list of questions used in both focus groups progressed from the general to the specific (Krueger & Casey, 2000). Group rapport was targeted using a general, simple and neutral opening question.

Clinical assessment forms:

1. What forms do you currently use?
2. What do you think are the strengths and weaknesses of the assessment instruments you have had experience using?

Possible follow up questions

- What aspects of current assessments would you keep?
 - What aspects of the current assessments would you change?
 - What is missing from the current assessments?
3. What would an ideal clinical assessment form look like?
 4. What are your thoughts about the draft Clinical Assessment of Physiotherapy Skills (CAPS) instrument?

3.3.1.7 Data management and analysis

Audio recording was used in the focus groups and two recorders were used in case one failed. On completion of the focus group, discussions were transcribed by a moderator. Within seven days of each focus group, the participants were sent a de-identified transcript of the group discussion. The participants were asked to confirm that the transcript accurately reflected the focus group discussion or request any amendments to the transcript. Once the focus group recordings had been transcribed and de-identified the tapes were erased.

The method of analysis chosen for this study was a synthesis of qualitative methods of thematic analysis. An a priori template of codes was developed based on theory and prior knowledge of issues relating to assessment of clinical performance. This was integrated with data driven codes (Boyatzis 1998, Crabtree and Miller, 1999, Fereday et al 2006).

The content analysis of transcripts and field notes was conducted independently by two reviewers and followed five steps:

1. Review of the transcript and observer notes
2. Creation of a coding guide using the focus group questions and existing knowledge of researchers of clinical education and assessment practice issues (Table 3.1)
3. Application of the coding guide to the focus group data
4. Identifying themes in the data, including those not covered by the questions
5. Interpreting the data (Barbour, 2005; Bogdan & Biklin, 1998; Hsieh & Shannon, 2005; Kitzinger, 1995; Melia, 1997; D. L. Morgan, 1988; Sandelowski, 2009).

There are no data that defend a 'best approach' to analysis of focus group data (Krueger & Casey, 2000). In this research identification of themes from focus group data was achieved

through the use of three techniques: word frequencies, key words in contexts, and a search for missing information. Researchers identified key words and then systematically searched the transcript to find all instances of the word or phrase. Each time a word was located, a copy of it and its immediate context were highlighted in the appropriate colour. Themes were identified by sorting the examples into lists with similar meaning. Rather than identifying emerging themes, the third technique involved searching for themes that appeared to be missing in the text. Once data were coded and collated under thematic headings, interpretation was assisted by asking the following four questions:

1. What was known and then confirmed or contested by the focus group data?
2. What was suspected and then confirmed or contested by the focus group data?
3. What was new that was not previously suspected?
4. What was suspected but not confirmed by the focus group data? (Krueger 1994)

3.4 Results (Part 1) Instrument development phase

Ethics approval was obtained from the Human Ethics Committees of Griffith and Monash Universities and from the Human Ethics Committees of each university where a physiotherapy program leader had agreed to participate in data collection in either the pilot trial or any of the subsequent field tests (see Appendix 3.4).

3.4.1 Construct Mapping

Figure 3.1 presents the first draft of a construct map that was developed by the research team to initiate discussion with stakeholders and inform the first iteration of instrument development. The right side of the construct map provided the foundation from which the proposed scoring system would be developed.

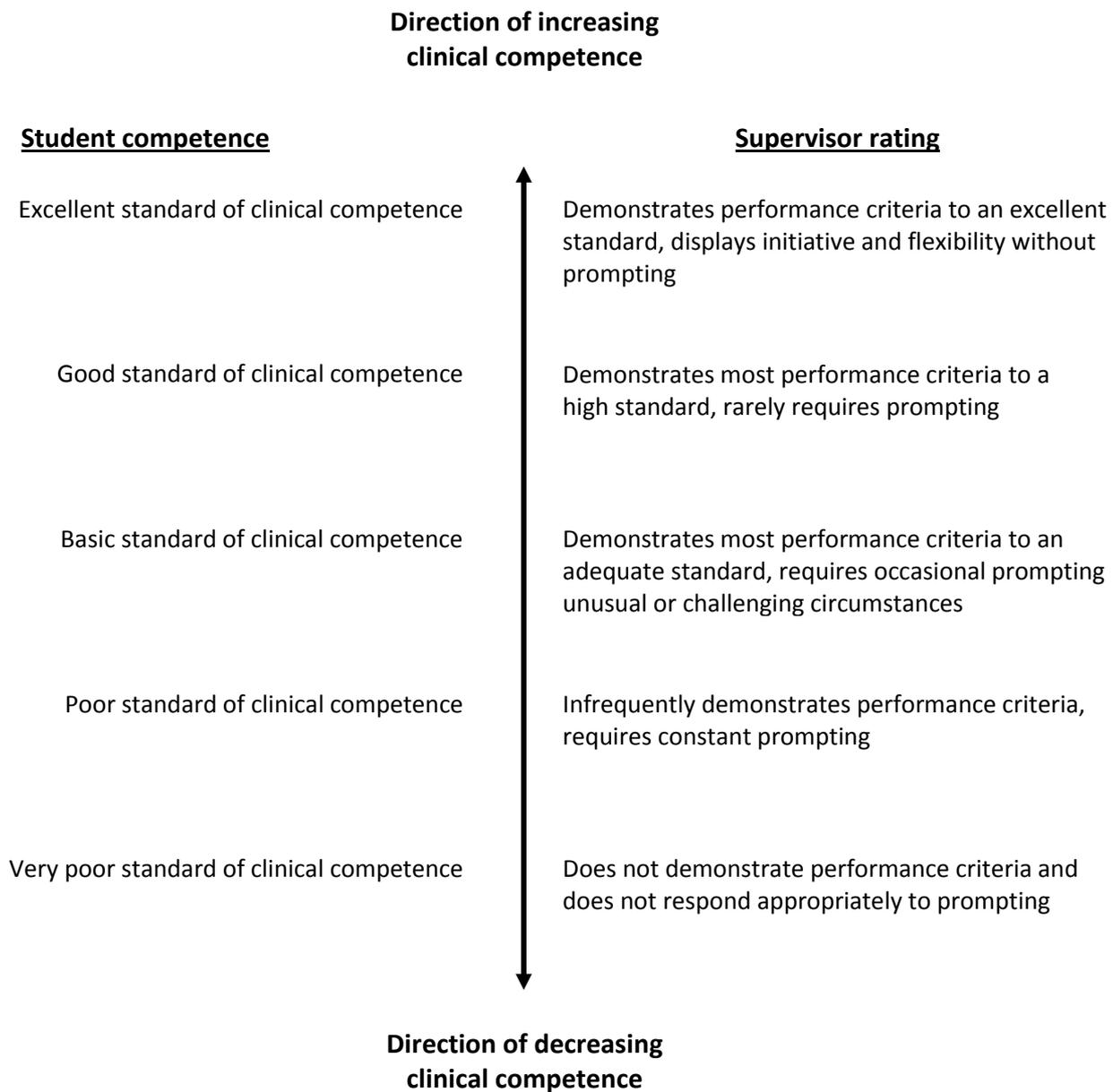


Figure 3.1: Construct map

3.4.2 Item design

Overall 634 potentially relevant items were assembled and entered into a spreadsheet. Construct domains were then determined based on observed relationships between items, domains identified by the Australian Physiotherapy Council (2006) and those used in current assessment instruments. Eight domains were identified. These were labelled communication, professional behaviour, assessment, analysis, planning, intervention, evidence based practice and risk management or safety. All identified items were able to be sensibly assembled under one of the eight domains. Following removal of duplicates and reduction of the item pool using the predetermined criteria, 18 items remained. These items were presented to the first two focus groups for discussion and refinement.

3.4.3 Outcomes space

The research group considered a range of scoring systems, the construct map and the literature underpinning rating scale development (refer to section 2.2.3) and initially chose a labelled five level rating scale (1 – 5) for each item (a unipolar adjectival scale) where 3 was considered a passing standard for the item. The research group was cognisant of the fact that the rating scale selected provided a platform for discussion only, that it would be discussed at the upcoming focus groups and would likely be modified based on feedback from the stakeholders (qualitative data) and from the results of Rasch analysis (quantitative data) that would be conducted on data collected during the pilot and field tests. The initial decisions on the format of the rating scale are summarised in Table 3.1. Additionally, the research group developed a list of performance indicators (PIs) for each item in the initial instrument. The PIs were presented to the first two focus groups for discussion and refinement.

A foundation instrument with 18 items (tentatively called the CAPS: Clinical Assessment of Physiotherapy Skills) was developed by the research group. The proposed items, scoring system and performance indicators were presented to the first two focus groups for consideration and recommendations (see Appendix 3.5 Initial draft of CAPS instrument (version 1)).

Table 3.1: Decisions to be made in relation to construction of a rating scale

Decision	Information on which decision was based (refer to Chapter Two for full details)	What was done
<p>Item score interpretation: criterion or norm referencing</p> <p>Level of measurement: categorical or continuous?</p> <p>Level of measurement: interval or ordinal</p>	<ul style="list-style-type: none"> • Since the purpose of work-based assessment is to judge the student's ability to demonstrate professional competencies in relation to specific standards rather than in relation to the performance of other students, the assessment was considered criterion referenced • Categorical judgements should not be used when the variable being measured can be assessed across a spectrum of ability • The continuous variable of performance demonstrated by a physiotherapy student is not interval level data 	<ul style="list-style-type: none"> • Use of a fixed minimum entry level standard presented as option for consideration • Competency of performance of physiotherapy students was considered a continuous variable as proposed in the construct map (Figure 3.1). • Rasch analysis chosen as the measurement model (refer to section 3.3.4) to enable conversion of ordinal data to interval data
Scale format	<ul style="list-style-type: none"> • The physical format of the rating scale may not be as critical as other factors. Kingstrom and Bass (1981) reviewed research comparing scale formats including VAS, Likert styles, or mixed formats. They concluded that there was little or no difference between the psychometric characteristics of the different scale formats • When considering assessment of professional competence, a unipolar scale is often the scale of choice as its structure more closely represents the underlying continuum of performance from very poor (incompetent) through to very high levels of competence with individual students demonstrating more or less of the variable (Wilson, 2005) 	<ul style="list-style-type: none"> • A labelled five level rating scale (1 – 5) for each item (a unipolar adjectival scale) where 3 was considered a passing standard for the item was the preferred scale format.
Number of response options for each item	<ul style="list-style-type: none"> • Reliability increases steadily up to 7 scale points, beyond which no substantial increases occur, even when the number of scale points is increased to as many as 100 (Cicchetti, et al., 1985) • Too few levels on the scale and information is lost and too many and the assessors cannot discriminate between the scale levels. Current research recommends between 5-7 (Streiner & Norman, 2003). Miller (1956) made the recommendation of 7 categories plus or minus 2. People tend to reduce a 100 point scale down to functional categories of around 5-7 (Cicchetti, et al., 1985; Fay & Latham, 1982; Hrachovy, et al., 2000; Munshi, 1990; Preston & Coleman, 2000; Streiner & Norman, 2003). 	<ul style="list-style-type: none"> • 5 point scale (0 – 4)
Even or odd number of response options	<ul style="list-style-type: none"> • Even number of categories in bi-directional scales forces the respondent to make a choice 	<ul style="list-style-type: none"> • Odd numbered scale with passing level as the mid point

Decision	Information on which decision was based (refer to Chapter Two for full details)	What was done
Location of a neutral point in the scale	<ul style="list-style-type: none"> Overall an odd or even number of categories is of no consequence in unipolar scales and should be decided on in relation to the population using the scale and their needs/preferences. Unbalanced scales (neutral is not at the mid point) can produce bias allowing a scale to produce more positive data than is accurate Respondents can find unbalanced scales frustrating to use, however they can be useful if there is likely to be an overwhelming response in a specific direction There is no absolute rule and the decision is best based on the needs of the instrument where it may or may not be appropriate to allocate a neutral point (Streiner and Norman 2003) 	<ul style="list-style-type: none"> An unbalanced scale proposed with three pass levels and two fail levels. The mid point representing passing level with 2 levels of fail below and 2 levels of superior performance above.
Labelling item response options	<ul style="list-style-type: none"> Most research shows there is little difference between scales with all points labelled and only the end points labelled (Dixon, Bobo, & Stevick, 1984). However (Dickinson & Zellinger, 1980) reported that respondents preferred rating scales with more verbal labels Using fully labelled scales may provide greater inter-rater reliability (Christian, et al., 2009). If only some boxes or points on the scale are labelled there is a tendency for these points to be chosen more often than unlabelled points (Streiner and Norman 2003) 	<ul style="list-style-type: none"> All points on the 5 point scale were labelled. Levels on the scale were labelled to imply increasing levels of competency Wording and scale design developed in conjunction with input from a broad cross section of end point users from the physiotherapy profession.
Adding a global rating scale	<ul style="list-style-type: none"> A combined or hybrid approach utilising both scores for specific items and global rating scales has been proposed to be the most suitable method of evaluating undergraduate clinical competence of medical students as it provides additional information regarding overall perception of ability that may not be captured by item scores (McIlroy, et al., 2002; P. J. Morgan, et al., 2001). Results of perceived overall performance can provide insight into the effects of tallying item scores to provide an overall grade for performance, as alignment between total item scores and global rating is desirable Analytic approach provides specific information on each aspect of performance whereas holistic scoring procedures may be preferable when an overall judgement of performance is required. 	<ul style="list-style-type: none"> Initially individual item based (analytic) scoring was chosen as this was similar to current practice.
Performance indicators	<ul style="list-style-type: none"> Loomis (1985b) reviewed the medical education literature and concluded that to improve rater reliabilities, competencies and associated 	<ul style="list-style-type: none"> For each item, a non-exhaustive list of performance indicators (PIs) were developed

Decision	Information on which decision was based (refer to Chapter Two for full details)	What was done
	<p>performance, standards need to be well defined in terms of observable behaviours or standards that describe the levels of mastery of the competency.</p> <ul style="list-style-type: none"> <li data-bbox="450 363 1234 480">• American Educational Research Association (1999) recommend that assessors are provided with examples of behaviours on which to base judgements of student ability to perform to the expected level of professional competence. 	<ul style="list-style-type: none"> <li data-bbox="1317 261 1883 480">• PIs were designed to provide examples of behaviours that were observable, measurable, able to be applied to guide improvement in performance, and transparent to student, educators and other stakeholders. Wording and examples were refined.

3.5 Results (Part 2) Consultation phase (focus groups)

Focus group one consisted of six participants who were all academics (n = 3) or clinical supervisors and/or clinical managers (n = 3) associated with the Bachelor of Physiotherapy at Monash University. In addition two members of the research group were also present (see Appendix 3.2).

Focus group two involved 18 participants, 14 university clinical education managers and four heads of physiotherapy programs at Otago University, The University of Melbourne, James Cook University and The University of Queensland (see Appendix 3.2). Participants of both focus groups confirmed that the transcripts accurately reflected the discussion and no requests or amendments to transcripts were made.

3.5.1 Content analysis

The research group investigated stakeholder views of item (content, wording and clarity of intent), scale (size, format and wording; pass level, and perception of likely use of the whole range of the scale), performance indicators (perceived utility, content, wording and clarity of intent), layout of the instrument (perceived utility, suggestions for improvement), and training in the use of the instrument.

After reviewing the transcript and observer notes, a coding guide was created using the focus group questions and a priori knowledge of clinical education and assessment practice issues (Table 3.2). Application of the coding guide to the focus group data enabled identification of themes and assisted interpretation of the data.

Table 3.2: Coding guide for content analysis of focus groups pre pilot trial

Code	Content
It & Dom	Items & domains of practice: content, wording and clarity of intent
Sc	Scale: size, format, wording
Pass	Pass standard: passing performance,
PIs	Performance Indicators : perceived utility, number, content, clarity of intent, wording, suggestions for additional PIs
IFor	Instrument format: layout of instrument, perceived utility, suggestions for improvement
Tr	Training in the use of the instrument: requirements of a training package
Other	Other key words, ideas, themes

The first analysis of the focus group transcripts identified that items, scale, pass level and performance indicators were the key words most frequently identified. Training and format of the instrument were less frequently mentioned. The issues of passing and passing standard were not only frequently mentioned but often with high levels of emotion identifying this as an issue of significance. To further inform instrument development key words related to concepts were categorised and considered in the context of issues relevant to instrument development. The process also aided identification of areas of consensus and disagreement in each focus group. The findings of the data from the first and second focus groups are summarised in Table 3.3.

3.5.2 Revised instrument (version 2) for pilot trial

Following consideration of the content analysis from the focus groups and relating the points raised with respect to the literature on instrument development, a number of changes were made to the CAPS instrument (version 1). A summary of the modifications is presented in Table 3.4. The amended 20-item APP instrument (version 2) used in the pilot trial is provided in Appendix 3.6.

Table 3.3: Summary of results of focus groups 1 and 2.

Target issue	Focus group results	Outcome / Actions arising
Items and domains of practice	<ul style="list-style-type: none"> Majority of participants agreed that additional items were required to cover the areas of identification and prioritisation of a client's problems and goal setting 	<ul style="list-style-type: none"> Consensus: 18 items increased to 20 Item 11 added: identifies and prioritises patient's/client's problems Item 12 added: Sets realistic short and long term goals with the patient/client.
Rating scale	<ul style="list-style-type: none"> Half of the participants considered need for additional scoring categories ie., 6 or 7 categories rather than 5 Clear definitions needed for each rating level Instrument needs to be useful to provide formative feedback to students as well as summative assessment of performance Student needs to be engaged in the formative assessment process Preference for 0 – 4 rather than 1 – 5 '1' for a failing performance sounds positive 	<ul style="list-style-type: none"> After discussion of recommendations in the literature, agreement to trial 5 point scale in pilot trial before making changes to rating scale Definitions for rating categories to be developed Agreement to trial 0 – 4 rating scale in pilot trial before making any changes to the rating scale Strategy to be developed to enable formative feedback eg half way formative feedback Assessors to be educated to use performance indicators to provide feedback on desirable performance Review following outcomes of pilot trial
Pass level	<ul style="list-style-type: none"> Majority of participants considered use of a novice (entry level) performance as pass standard for items was appropriate but could cause difficulty for some educators who usually used to an alternative model of grading students against 'the expected competency during the first practice block in third year' or 'the expected competency during the last practice block in fourth year' 	
Performance Indicators	<ul style="list-style-type: none"> All participants agreed clear behavioural examples to guide ratings were essential All participants considered PIs would be useful to guide assessment Some participants thought specific PIs were needed to cover clinical areas such as community health and paediatrics 	<ul style="list-style-type: none"> Consensus on need for performance indicators Research group to modify wording on specific PIs to ensure comprehensive, clear, and appropriate for all clinical areas.
Layout of instrument	<ul style="list-style-type: none"> All participants wanted instrument to be contained to one page Name change from CAPS requested because the word 'clinical' was considered to be too 'hospital based' and not adequately inclusive of areas of physiotherapy practice 	<ul style="list-style-type: none"> Consensus to aim for one page layout Consensus for a name change of the instrument from Clinical Assessment of Physiotherapy Skills (CAPS) to Assessment of Physiotherapy Practice (APP)
Training	<ul style="list-style-type: none"> Important to ensure consistency of ratings across placements Training must be easy to access by all educators Training must be time efficient 	<ul style="list-style-type: none"> Consensus in regard to need for training to be consistent and time efficient

Table 3.4: Summary changes to CAPS (version 1) prior to pilot trial

Requested modifications to APP (v2)	CAPS (version 1)	APP (version 2) Pilot trial
Instrument name change Rewording of Item 10 Addition of 2 items	<i>Clinical Assessment of Physiotherapy Skills</i> <i>Sensibly</i> interprets assessment findings 18 items	<i>Assessment of Physiotherapy Practice</i> <i>Appropriately</i> interprets assessment findings 20 items • <i>Item 11</i> added: identifies and prioritises patient's/client's problems • <i>Item 12</i> added: Sets realistic short and long term goals with the patient/client
Change scoring scale from 1 – 5 to 0 – 4 plus not assessed	See below	See below
Change wording on the scoring scale descriptors	<ol style="list-style-type: none"> 1. Does not demonstrate performance criteria and does not respond appropriately to prompting 2. Infrequently demonstrates performance criteria, requires constant prompting 3. Demonstrates most performance criteria to an adequate standard, requires occasional prompting in unusual or challenging circumstances 4. Demonstrates most performance criteria to a high standard, rarely requires prompting 5. Demonstrates performance criteria to an excellent standard, displays initiative and flexibility without prompting 	<p>0 = Infrequently demonstrates performance indicators, requires constant prompting with usual/typical (non-complex) patient presentations</p> <p>1 = Demonstrates some performance indicators to an adequate standard, requires frequent prompting with usual/typical (non-complex) patient presentations</p> <p>2 = Demonstrates most performance indicators to an adequate standard, requires prompting in atypical or complex patient presentations</p> <p>3 = Demonstrates most performance indicators to a high standard, requires occasional prompting</p> <p>4 = Demonstrates most performance indicators to an excellent standard, rarely requires prompting</p>
Change order of domains of practice with Professional behavior before communication	<i>Communication; professional behaviour; assessment; analysis and planning; intervention; evidence based practice; risk management</i>	<i>Professional behaviour; communication; assessment; analysis and planning; intervention; evidence based practice; risk management</i>

Requested modifications to APP (v2)	CAPS (version 1)	APP (version 2) Pilot trial
Modify wording on PIs to ensure comprehensive, clear, and appropriate for all clinical areas.	<i>Patient</i>	<p data-bbox="1218 264 1626 288"><i>Patient/client/carers (as appropriate)</i></p> <p data-bbox="1218 328 1552 352">PIs for 2 new items developed</p> <p data-bbox="1218 360 1877 384">Item 11: <i>Identifies and prioritises patient’s/client’s problems</i></p> <ul data-bbox="1218 392 1910 488" style="list-style-type: none"> • <i>generates a list of problems from the assessment</i> • <i>collaborate with the patient/client to prioritise their problems</i> • <i>considers patient’s/client’s values, priorities and needs</i> <p data-bbox="1218 528 1834 584">Item 12: <i>Sets realistic short and long term goals with the patient/client</i></p> <ul data-bbox="1218 592 1960 791" style="list-style-type: none"> • <i>negotiates realistic short and long term treatment goals in partnership with patient/client</i> • <i>formulates goals that are specific, measureable, achievable, relevant and timely</i> • <i>considers physical, emotional and financial costs and relates them to likely gains of physiotherapy intervention</i> <p data-bbox="443 823 1178 887">1. Communicates effectively and appropriately – verbal/non-verbal <i>Greets patients/clients and carers appropriately</i></p> <p data-bbox="443 919 1189 1046">14. Monitors the effect of intervention</p> <ul data-bbox="443 959 1189 1046" style="list-style-type: none"> • <i>Monitors patient throughout the intervention and makes modifications as appropriate. Monitors and analyses relevant health indicators appropriately</i> <p data-bbox="443 1094 999 1118">17. Applies evidence based practice in patient care</p> <p data-bbox="1218 823 1951 887">5. Communicates effectively and appropriately – verbal/non-verbal</p> <ul data-bbox="1218 863 1552 887" style="list-style-type: none"> • <i>Greets others appropriately</i> <p data-bbox="1218 919 1653 943">16. Monitors the effects of intervention</p> <ul data-bbox="1218 959 1924 1054" style="list-style-type: none"> • <i>Monitors patient throughout the intervention and makes modifications as appropriate.</i> • <i>Monitors and analyses relevant health indicators appropriately</i> <p data-bbox="1218 1094 1951 1222">19 Applies evidence based practice in patient care (PI added in)</p> <ul data-bbox="1218 1126 1951 1222" style="list-style-type: none"> • <i>Options for physiotherapy intervention are indentified and justified, based on the needs of the patient/client, on best evidence and available resources</i>

(Code: amendments to instrument highlighted in *italics*)

3.5.3 Mapping the standards

Following feedback from focus groups and prior to the pilot trial, a detailed mapping of the item content to the Australian Physiotherapy Council Standards (2006) was completed (Appendix 3.7). This provided validity evidence based on test content and ensured that there was no construct underrepresentation or construct irrelevance within the items on the instrument (American Educational Research Association, 1999; Cook & Beckman, 2006; Goodwin, 2002; Mokkink, et al., 2010).

3.6 Discussion

In the early stages of instrument development the research group was aware that the initial version of the instrument would undergo a continuous cycle of testing and refinement based on quantitative and qualitative analysis. Refinements to the instrument prior to the pilot trial were informed by the qualitative data from focus groups, review of literature and evaluation of relevant instruments and documents. Stakeholder support was critical to achieve the dual aims of a psychometrically sound instrument with ownership and uptake by the profession.

The focus groups resulted in the number of items in the pilot version increasing to 20, spread across eight domains of practice. Issues raised in focus groups, such as the number of categories on the rating scale, were targeted for evaluation following pilot testing. Participants agreed that they were unsure of the merits of changing the scale to include seven or more rating categories. Participants in both focus groups agreed to wait on the outcome of the pilot trial before revisiting the number of scoring categories. Participants agreed that a 0 - 4 scale was as acceptable as a 1 – 5 scale, that end scale aversion was unlikely and that zero provided clearer feedback on the level of performance than a rating of 1. There was consensus that the instrument must be able to be used to provide both formative and summative assessment, that the students needed to be engaged in self assessment and that formative feedback sessions should include the development of strategies to improve clearly defined performance targets.

Wass et al (2001b) recommended that the minimum appropriate standard be decided upon prior to use of an assessment instrument. Since a clear minimum standard at which a student was deemed competent was required when assessing professional competence, norm referencing (i.e. relative to peer performance) was considered inappropriate. Criterion referenced testing of aspects of professional competence targeted by each item set the minimum standard of performance at a level considered evidence of fitness to practice according to qualified practitioners.

The decision regarding how to set a pass standard for items, and for overall assessment, generated a great deal of discussion at both focus groups. Some university programs have traditionally used entry-level competence as the benchmark against which to judge student performance, while others have used the performance that would be expected at the particular stage of the course (e.g., second year standard, third year standard). An advantage of marking students against acceptable entry level standards is that, theoretically at least, all assessors could assess against a set standard. In discussions about entry level/beginning physiotherapist standards there was clear consensus from participants that for consistent use of an instrument across programs, students should be judged on each item against the minimum performance targets expected of a novice (entry-level) practitioner. The alternative model of grading students against 'the expected competency during the first practice block in third year' or 'the expected competency during the last practice block in fourth year' created individually constructed and unregulated assessment targets, and limited the opportunity for discussion regarding what that standard should look like and how to best support all students to achieve desirable standards of performance. The focus group participants agreed that many students had only one clinical block within which to gain skills in core areas of practice e.g., neurological rehabilitation. It was therefore essential that the pass standard at the end of that block was entry level practice. The target of clinical education was the acquisition of a minimum acceptable level of skills irrespective of when each clinical unit was completed. A target of entry level competence enabled ranking of students relative to a common standard.

Participants agreed that the behavioural performance indicators for each item were valuable. Approximately a third of participants suggested that the performance indicators

be reviewed to encompass all areas of physiotherapy practice, in particular, paediatric rehabilitation and community health. The research group agreed to address this issue prior to pilot testing (refer to Table 3.4 for more detail). Participants were also advised that the PIs provided were not an exhaustive list expected to cover all situations in which students might practice and that they could develop context specific PIs in collaboration with the student.

The one page layout of the instrument was considered user-friendly. A change of name for the instrument (CAPS) was requested as the word 'clinical' was considered to be too 'hospital based' and not representative of all areas of physiotherapy practice and the term 'skills' was thought to reduce the holistic practice of a physiotherapist to a list of technical skills. After considerable discussion and debate, Assessment of Physiotherapy Practice (APP) was agreed upon.

Training in the use of instrument was considered desirable to ensure familiarity with the items, indicators and rating scale. It was agreed that training would have to be accessible and time efficient if educators were to participate. This topic was scheduled for follow up in future focus groups.

The revised instrument (version 2), scoring system and performance indicators are presented in Appendix 3.6. It was this version of the instrument that was trialled in the first pilot study, which will now be described in Chapters 4 and 5.

4. Chapter Four: Pilot Trial - Quantitative evaluation

4.1 Introduction

This chapter of the thesis reports the first pilot test of the APP, an instrument to assess professional competence in physiotherapy students. This pilot trial was designed to investigate the nature of the scores when the instrument was used by clinical educators (graduate physiotherapists) to assess directly observed performance of undergraduate physiotherapy students in the authentic practice environment. It was of particular interest to assess the behaviour of scores for different items, and to determine whether item scores provided evidence of measurements of a single underlying construct.

Data were collected and analysed using the Rasch Measurement Model (section 2.2.4) because it provides a sophisticated method for scale development. The pilot trial also provided an opportunity for development of standardised user guidelines, refinement to the instrument prior to field testing and formed part of the iterative action research nature of the study illustrated in Chapter Two, Figure 2.5.

Qualitative and quantitative data on instrument performance were applied in assessing the validity and reliability of the measurements and feasibility of its use in the practice environment. Raw scores (0-4) for each of the twenty items on the pilot version (version 2) of the APP (Appendix 3.6) and the sum of item scores (a total score) for each student assessed during the pilot trial were examined. Feedback from clinical educators who used the instrument during the pilot trials was collected through focus groups conducted after pilot testing and prior to commencement of field testing. The results of the focus group analysis are reported in the following chapter (Chapter Five).

4.2. Method

4.2.1 Participants – students and clinical educators

Students

Participants in the pilot trial were students enrolled in a 4-year Bachelor of Physiotherapy program at La Trobe University, Victoria, Australia. The Bachelor program was accredited by the national accrediting body, the Australian Physiotherapy Council (APC). The pilot version

of the APP (version 2) (Appendix 3.6) was used to assess students during usual 5-week clinical placements blocks scheduled in one university semester in 2006. Students attended clinical placements on a full-time basis (32-40 hours/week).

Clinical educators

During clinical placements, students were supervised by clinical educators (registered physiotherapists) in 1:1 to 1:4 educator:student ratios. Recruitment was targeted to achieve representation of educators by location (metropolitan, regional/rural and remote), clinical area of practice, years of experience as a clinical educator/supervisor or manager, organization (private, public, hospital based, community based and non-government). While all educators were employed in facilities spread across one Australian state, Victoria, there were no *a priori* reasons to consider that this group of educators would be different from those working in other Australian states.

The number of completed student assessments suitable for informative Rasch analysis was determined based on recommendations by (Linacre, 1994). Linacre recommends a sample size of 100 (n range, 64 – 144) to provide 95% confidence within +/-0.5 logits or 150 (n range, 108-243) to provide 99% confidence within +/-0.5 logits for item calibration. If a test is not well targeted, a larger sample size (243) is recommended. In the current study a minimum sample size of 243 APP results was considered feasible, enabling adequate precision regardless of the targeting of the group or the distribution across the response options of each item (Linacre, 1994). Independence of participants is not a requirement for Rasch analysis as the relationship of interest, between item and total scores, is considered independent under repeated assessments (Bond and Fox 2007). Assessments from more than one placement for some participants were therefore included.

4.2.2 Pilot trial testing procedure – prior to commencement of clinical unit

Information about the intended research was provided to students and educators involved in the clinical placements and their consent was sought (Appendix 4.1). Student's assessment data were excluded if either the student or their clinical educator did not consent to participate in the research. Participants were advised that all data would be de-

identified prior to analysis. As the APP was used in addition to usual grading procedures, students were also informed that the scores provided by the clinical educators based on the APP instrument would not be used in establishing grades for the clinical unit.

4.2.2.1 Instructions to clinical educators

As clinical educators were located in facilities spread across Victoria, face to face training for all educators in the pilot trial was impractical. The financial and time costs for educators to attend formal training sessions were not justifiable at a stage in instrument development when substantial refinement of the instrument during field testing was likely. To standardise use of the instrument, all educators were provided with written instructions on the APP (Appendix 4.2) and how to complete and return forms. A member of the research group (MDav) was available to answer questions by phone or email. The pilot trial also provided an opportunity to obtain feedback from clinical educators on problems encountered in using the instrument, issues requiring clarification, reflections on the instrument utility and feasibility, and the kind of training programs (amount, timing and location) that would be helpful prior to subsequent field testing. Consenting students were provided with a copy of the instrument and information describing the research. (Appendix 4.1)

4.2.3 Pilot trial testing procedure – during the clinical unit

Clinical educators were advised to conduct the clinical unit according to normal procedure. This meant providing formative feedback to students midway through the unit using the standard La Trobe University forms. The APP instrument was not completed at mid unit. On completion of the clinical placement, the students were assessed on their performance by their primary clinical educator using two instruments, the APP and La Trobe University's clinical assessment instrument in use at the time of the pilot trial. Educators were instructed to assess the student's performance using the APP at the end of the clinical unit prior to completing the required university assessment documents. If a student had more than one clinical educator during the placement, only one educator was required to return a completed APP instrument. Students did not view the completed APP instrument.

4.2.4 Pilot trial testing procedure – Data management and analysis on completion of the clinical unit

Completed forms were returned to one of the researchers (MDav) at La Trobe University by mail, entered into Microsoft Excel 2003 and de-identified. The aim of data analysis was to investigate properties of the scores obtained using the instrument. Data analyses were performed using SPSS 14.0 (SPSS Inc.) and RUMM2020 software for Rasch analysis (Andrich, Lyne, Sheridan, & Luo, 2003). Data were coded as missing if an item was not scored on the 0 – 4 rating scale.

4.2.5 Rasch analysis

Rasch analysis is the formal testing of an instrument against a mathematical measurement model developed by the Danish mathematician Georg Rasch (1960). The model specifies what should be an expected pattern of responses to items if measurement (at interval level) is to be achieved. Data from rating scales or questionnaires that are summed into a total score are tested against what the Rasch Measurement Model determines to be the ideal response pattern if interval level measurement is to be achieved; this is based on a probabilistic form of Guttman scaling (Andrich, 1988; Tennant, McKenna, & Hagell, 2004). A variety of statistics are applied to determine if the data have an adequate fit to the model, and these are discussed later.

Rasch analysis transforms ordinal data to log odds ratios (logits) to place persons (ability) and items (difficulty) on the same interval level scale (Andrich, 2004). In this logit transformation, items can be ranked by difficulty based on the typical scores for the item and students can be ranked on ability based on total scores across items. If a student's ability to perform a particular activity is lower than the required ability for that particular task, the probability of the student achieving a low score rating is high. Conversely if a student's level of ability is greater than the ability required for a particular task, the probability is high that the student will achieve a high score rating (Andrich, 1978; Bond & Fox, 2007; Lamoureux, Pallant, Pesudovs, & Hassell, 2006).

Dichotomous (Rasch, 1960) and polytomous (Andrich, 1978) versions of the Rasch model are available. For polytomous data there are two options: the rating scale model (Andrich,

1978) or the partial credit model (G. Masters, 1982). The rating scale model expects the distance between the thresholds across items to be equal. A threshold is the probabilistic midpoint (50/50) of a score falling between any two adjacent categories (e.g. being classified as a pass on the item or good on the item). This means that the metric distance between the thresholds separating categories one and two is the same as that separating categories two and three across all items (Tennant & Conaghan, 2007). The alternative partial credit model allows the thresholds to vary for each individual item. RUMM2020 software enables the most appropriate model to be chosen by initially conducting a Likelihood Ratio test that evaluates score characteristics and identifies the most suitable model to apply. If the obtained probability that thresholds are not uniform across items is significant ($p < .05$) then the partial credit model is more appropriate. Rasch analysis enables formal assessment of the measurement properties of an instrument through investigation of overall model fit, overall person fit and item fit, individual item fit, thresholds, targeting, person separation index (PSI), differential item functioning (DIF) and local independence (dimensionality) (J. F. Pallant & Tennant, 2007).

4.2.5.1 Overall model item and person fit

To examine the hypothesis that items on the instrument assessed a single underlying construct, the difference between the observed item response and that expected if the data fit a Rasch model was investigated. Three overall fit statistics were considered. Firstly the overall fit of the responses to the model was described by an item-trait Interaction statistic reported as a chi-square. A non-significant Chi-Square Item-Trait Interaction statistic would indicate no significant variation from the predicted model and provide evidence of trait invariance, that is, that the hierarchical ordering of the items is consistent across all levels of the underlying trait (defined as professional competence). Additionally, two item-person interaction fit statistics were calculated. Rasch analysis conducted by RUMM2020 converts the item-person fit statistics to an approximate z-score. If the items and persons fit the model, the mean item or person fit residual would be approximately 0 with a standard deviation (*SD*) around 1. If the *SD* value is above 1.5 this would suggest a problem (J. F. Pallant & Tennant, 2007).

4.2.5.2 Individual Item and Person Fit

In addition to overall summary fit statistics, individual item and person fit statistics were assessed using both residuals and a chi-squared statistic. The fit residuals were expected to be within the range of -2.5 to +2.5 if the data fit the model. Misfitting items or persons are indicated by two statistics: a fit residual value beyond ± 2.5 or a significant chi-square probability value. Items with large negative residual values indicate a high level of predictability in responses and signal possible item redundancy. Items with large positive residual values suggest an item does not contribute to the measurement of a unidimensional construct (G. N. Masters & Keeves, 1999).

4.2.5.3 Threshold ordering of polytomous items

Rasch analysis enables examination of whether or not the category ordering of the polytomous items behave as anticipated. For a good fit to the model it is expected that for any item respondents with high levels of the attribute (professional competence) would be scored higher than individuals with low levels of the attribute. In Rasch analysis this is demonstrated by an ordered set of response thresholds for each item. The term threshold refers to the point between two categories where either response is equally probable, that is, for example, the point between category 1 and 2 where there is a 50/50 probability of scoring in either category. The APP has five rating options of increasing levels of professional competence and therefore four thresholds for each item. Each threshold has a location on the logit scale and each item has an average location. Ordered thresholds indicate that the respondents (clinical educators) are able to use the response categories (scoring scale) in a manner consistent with the level of the trait (competency) being measured. This occurs when the educators have no difficulty discriminating between response options. Rasch analysis allows identification of whether the steps in the rating scale attract this expected pattern of responses or whether the thresholds are disordered, that is, when the probability of selecting each level does not occur in the manner predicted.

4.2.5.4 Targeting

It is important, particularly in clinical practice, that the assessment items are appropriately targeted for the population being assessed i.e. they are neither too easy nor too hard.

Poorly targeted measures result in floor or ceiling effects, and this would mean that either very weak or very strong students may not be appropriately graded. Comparison of the persons mean fit residual score with that of zero set for the mean difficulty of the items provides an indication of how well targeted the items are for the people in the sample (J. F. Pallant & Tennant, 2007).

For a well targeted measure the mean location for the persons will also be around zero. If the mean value for the persons is positive this would indicate that the sample was at a higher level (professional competence) than the average of the scale. A negative mean value would indicate the opposite (Pallant and Tennant 2007). Targeting is also judged by visual examination of the spread of person and item thresholds on the logit scale. Poor targeting occurs when item thresholds are clustered at certain points along the logit scale leaving gaps, and where respondents have a higher or lower ability than the most or least difficult item threshold (Davidson & Keating, 2002; J. F. Pallant & Tennant, 2007).

4.2.5.5 Person separation index

The Person Separation Index (PSI) provides an indication of the internal consistency of the scale and the power of the APP to discriminate amongst respondents with different levels of the trait (professional competency) being measured. The PSI is the same as Cronbach's alpha with the logit value replacing the raw score in the same formulae. The higher the reliability of person separation, the more groups the instrument is able to detect. A reliability coefficient of 0.8 indicates that two groups can be identified, and 0.9 four or more groups (Wright & Masters, 1982).

4.2.5.6 Dimensionality

One of the primary tenets underpinning Rasch analysis is the concept of unidimensionality. All items on the scale should measure the same construct. Unidimensionality of the APP was tested using a principal components analysis of the Rasch model residuals (Smith, 2002). When the Rasch factor (underlying construct) has been removed, the residuals are what remain and it is expected that there will be only random associations between items. If a set of items are truly unidimensional, then if a person is rated using any subset of the item in

the instrument their ratings should provide the same person ability estimates as if they had been rated using the entire test. Two item subsets are created from items loading positively and negatively on the first residual factor in the PCA. Only those items with loadings greater or less than 0.3 were considered. The next step is to investigate if the person estimates (location values) based on scores that underpin each set of items are significantly different using independent t-tests. A confidence interval for a binomial test of proportions is calculated for the observed number of significant tests (Tennant & Pallant, 2006).

4.2.5.7 Differential item functioning (DIF)

In the Rasch measurement model, a scale should function consistently irrespective of the subgroups within the sample being assessed. For example male and female students with equal levels of the underlying construct being measured should not be scored significantly differently (Lai, Teresi, & Gershon, 2005; Teresi, 2001).

The primary purpose of the pilot trial was to enable preliminary investigation of the functioning of the items and scoring system of the APP instrument. Since the research group was aware that there would be at least two full field tests of the instrument, a decision was made to keep the additional work load required of the clinical educators involved in the pilot trial to a minimum, thus amplifying compliance of the stakeholders in the next phases of the research. Collection of the necessary demographic data in order to enable testing of DIF was delayed until field testing and is presented and discussed in Chapters 7 and 9.

4.3. Results

Ethics approval was obtained from the Human Ethics Committee of La Trobe, Griffith and Monash Universities (Appendix 3.4).

4.3.1 Participants' characteristics

Data were obtained for 295 completed assessments of 181 students. Students were undertaking either their first major clinical placement in third year (n=108 completed assessments) or final placements in fourth year (n= 187 completed assessments). All

placements were of five weeks duration. Students had a mean age of 22 ± 4 years and 66% were female. There were 132 clinical educators with a mean age of 35.2 ± 9 years and 70% were female. Multiple assessments (maximum 2 for any student) occurred for 38% of students.

4.3.2 Characteristics of item and instrument scoring

Table 4.1 presents the descriptive statistics of the raw score for each item, the total raw score for the 20 summed item scores and the frequencies of use of each rating scale category for the 20 items on the 295 completed assessments.

The mean (M) of the 108 completed assessments from the student's first major clinical placement in third year was 62.46 ($SD=4.51$) (raw score range 0 – 80 converted to 0 - 100) and for the final placements in fourth year (187) the mean was 72.15 ($SD= 9.87$).

Table 4.1: Descriptive statistics (n=295 completed assessments)

Item	N	Mean	Standard Error of Mean	Std. Dev	Rating 0		Rating 1		Rating 2		Rating 3		Rating 4	
					Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
1	294	3.65	.033	.562	0	0	1	0.3	10	3.4	79	26.8	204	69.2
2	295	3.63	.036	.613	0	0	1	0.3	18	6.1	69	23.4	207	70.2
3	295	3.76	.030	.513	0	0	1	0.3	9	3.1	49	16.6	236	80.0
4	293	3.50	.038	.644	0	0	2	0.7	18	6.1	104	35.3	169	57.3
5	295	3.35	.043	.732	0	0	3	1.0	36	12.2	110	37.3	146	49.5
6	295	3.40	.038	.652	0	0	2	0.7	21	7.1	128	43.4	144	48.8
7	295	3.27	.040	.686	1	0.3	2	0.7	28	9.5	149	50.5	115	39.0
8	293	3.19	.038	.643	0	0	2	0.7	32	10.8	168	56.9	91	30.8
9	295	3.11	.042	.728	1	0.3	6	2.0	40	13.6	162	54.9	86	29.2
10	295	3.01	.044	.751	1	0.3	7	2.4	55	18.6	158	53.6	74	25.1
11	295	3.10	.041	.702	0	0	4	1.4	47	15.9	159	53.9	85	28.8
12	295	3.05	.039	.673	0	0	7	2.4	38	12.9	182	61.7	68	23.1
13	294	3.21	.041	.697	0	0	4	1.4	35	11.9	151	51.2	104	35.3
14	295	3.28	.041	.710	1	0.3	4	1.4	26	8.8	143	48.5	121	41.0
15	295	3.27	.044	.748	1	0.3	4	1.4	35	11.9	129	43.7	126	42.7
16	295	3.20	.042	.727	0	0	3	1.0	45	15.3	137	46.4	110	37.3
17	295	3.13	.041	.703	0	0	2	0.7	50	16.9	151	51.2	92	31.2
18	288	3.07	.043	.729	1	0.3	3	1.0	52	17.6	151	51.2	81	27.5
19	293	3.22	.042	.726	1	0.3	2	0.7	40	13.6	139	47.1	111	37.6
20	294	3.44	.041	.707	0	0	3	1.0	23	7.8	108	36.6	160	54.2
Tot. score for 20 items	295	65.68	0.580	9.88	Range of total raw scores for 20 items: minimum=25; maximum=80									
		/80												

Legend: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies evidence based practice 20 = assesses risk; Std. Dev.= standard deviation; N=number

The data presented in Table 4.1 show that overall there is infrequent use of the ratings 0 and 1 which represent failure to reach minimum entry level standard of performance for an item. This is acceptable considering 187 of the 295 completed assessments were from fourth year students completing their final placement before graduation. The total score for the 20 items ranged from 25 – 80 highlighting an acceptable spread of scores. Similarly all 5 points on the 0 – 4 rating scale were used. Item 18 (undertakes discharge planning) was the item most frequently not scored. However this occurred only 7 times out of a possible 295 occasions. Of the possible 5900 item scores (295 x 20) there were 16 items not scored. The missing data rate was therefore 0.27%.

4.3.3 Rasch analysis: Model

The Likelihood Ratio test was significant ($p < 0.001$) so the partial credit model was used.

4.3.4 Rasch analysis: Overall Model Fit

The Chi-Square Item-Trait Interaction statistic was 82.28 ($df= 80, p= 0.40$) with a Bonferroni adjusted alpha (α) value of .0025 (.05/20). The chi-square probability value of 0.40 indicated excellent fit between the data and the model.

4.3.5 Overall Item and Person Fit

The residual mean value for items was -0.55 ($SD 1.10$), indicating no serious misfit of items to the model. Similarly the residual mean value for persons was -0.40 ($SD 1.35$) indicating no misfit among the respondents in the sample.

4.3.6 Individual Item and Person Fit

There were no positive item fit residuals above 1.46 (range 0.3 – 1.46). There were 3 people with positive fit residuals above +2.5. Investigation of these individual results revealed two instances of unexpected scoring on item 19 (evidence based practice) and one on item 18 (discharge planning).

4.3.7 Threshold ordering of polytomous items

There were no disordered thresholds for any of the 20 items in the pilot trial. The threshold map for the 20 items is illustrated in Figure 4.1. An additional example of ordered response thresholds is demonstrated in the category probability curves shown in Figure 4.2.

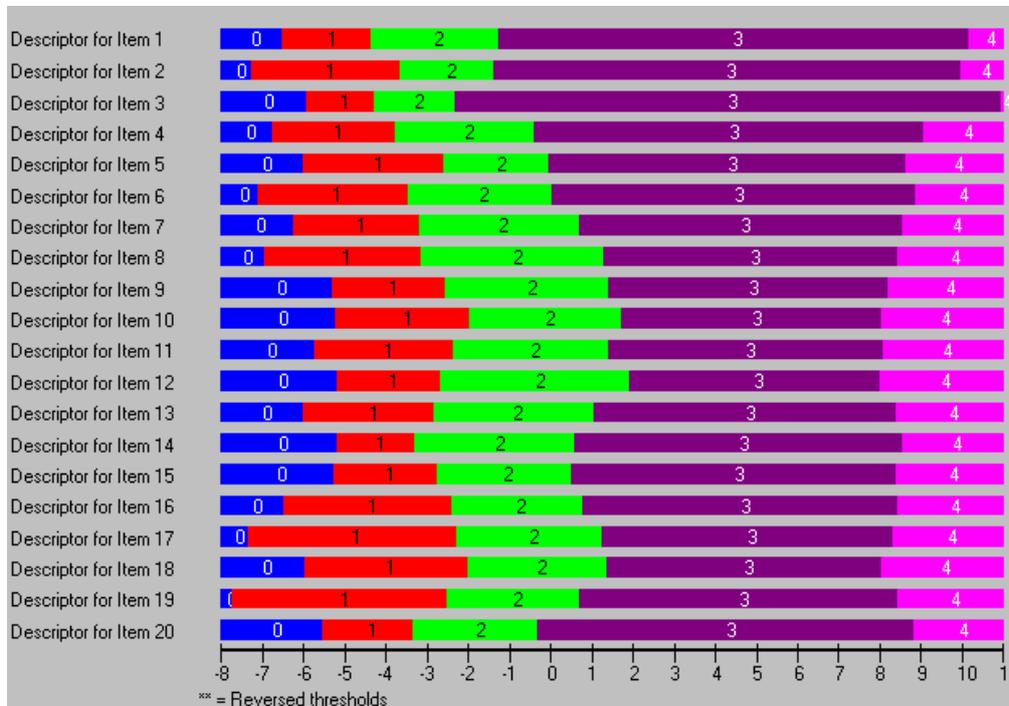


Figure 4.1: Threshold map for items 1 - 20

Legend: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies evidence based practice 20 = assesses risk

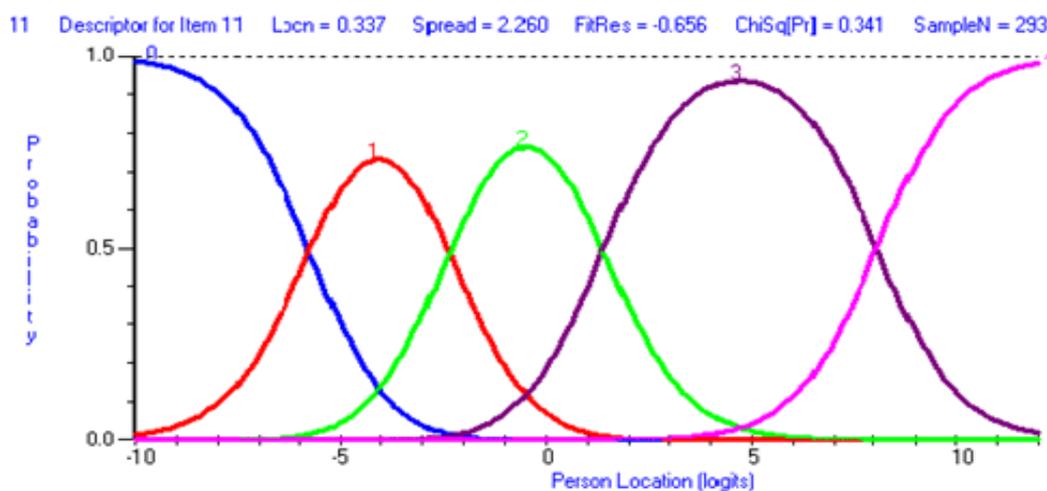


Figure 4.2: Category probability curves showing ordered thresholds for item 11 (Identifies and prioritises patient's/client's problems). Locn=location; FitRes= Fit Residual; ChiSq[Pr]=chi-square probability

4.3.8 Targeting

Figure 4.3 shows the distributions of the students (top half of the graph) and item thresholds (bottom half of the graph) for the APP total score on a logit scale. The fit residual mean value for persons was -0.40 (*SD* 1.35) (see section 4.3.5) indicating that the test was reasonably well targeted for use with this group of students.

Overall there appeared to be acceptable matching of item difficulty with person abilities. There was some mismatching at the left and right of the graphs (i.e. item thresholds that have no equivalent person abilities as all students could achieve these, and person abilities that have no equivalent item threshold difficulties that could differentiate their performance. At the far right hand end there are several item thresholds where the most able students had difficulty scoring well and there is a gap in item threshold distributions between location scores 2 - 8. These aspects of targeting require further investigation in the field testing phase.

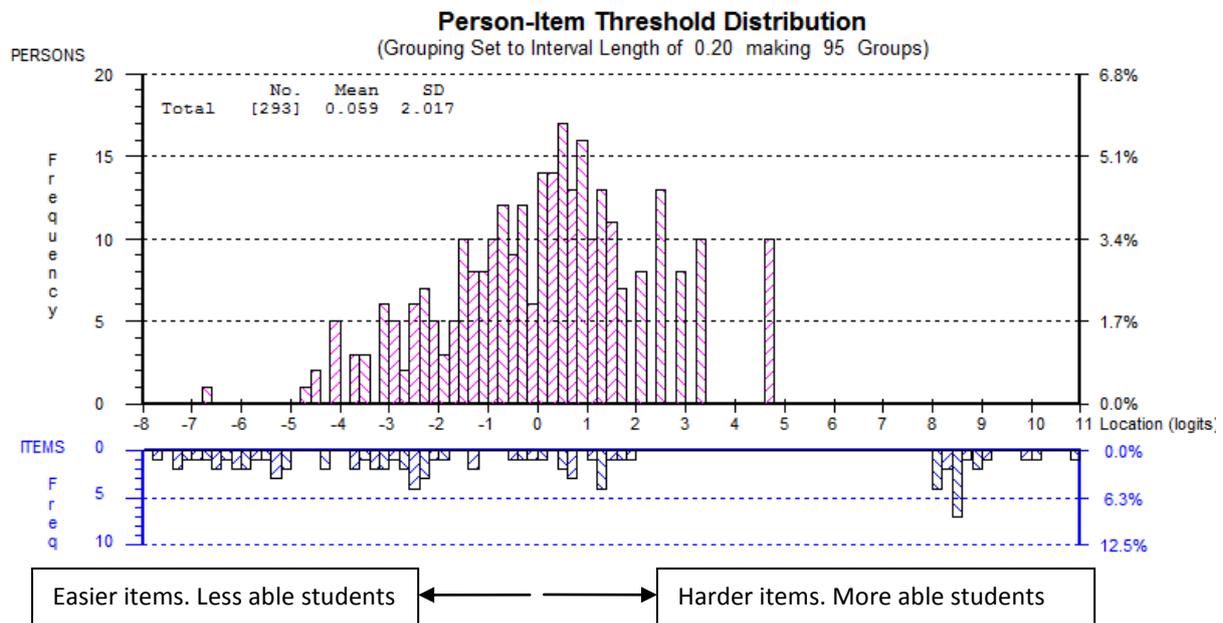


Figure 4.3: Person item distribution graph for total APP scale

4.3.9 Hierarchy of item difficulty

The sequence or hierarchy of difficulty of the 20 items on the APP are shown in Table 4.2.

The ranking is by the point estimate of the average logit location for each item.

Table 4.2: Item order, average location and standard error (SE) from least to most difficult of the 20 items

Item	Location	SE	FitResid	DF	ChiSq	DF	Prob
2	-0.583	0.135	-0.347	274.35	1.556	4	0.816624
1	-0.501	0.14	-0.654	273.41	2.264	4	0.687321
4	-0.466	0.129	0.74	272.48	3.884	4	0.422
6	-0.424	0.125	1.467	274.35	9.327	4	0.053435
3	-0.4	0.154	-0.62	274.35	2.455	4	0.652683
5	-0.274	0.118	0.329	272.48	2.386	4	0.665245
8	-0.09	0.127	-1.448	272.48	6.249	4	0.181313
20	-0.087	0.121	0.618	273.41	3.887	4	0.421451
7	-0.061	0.122	-0.293	274.35	4.81	4	0.307348
18	-0.02	0.118	-1.684	274.35	4.456	4	0.347754
19	0.001	0.116	0.535	274.35	3.149	4	0.533186
16	0.084	0.116	-1.905	274.35	3.281	4	0.511983
13	0.155	0.12	-2.048	273.41	6.713	4	0.151857
14	0.167	0.12	-1.183	274.35	4.636	4	0.326715
15	0.221	0.115	-0.059	274.35	2.121	4	0.713592
11	0.337	0.117	-0.656	274.35	4.51	4	0.341359
17	0.348	0.117	0.718	267.79	2.766	4	0.597671
9	0.433	0.118	-1.956	274.35	2.553	4	0.635141
12	0.512	0.123	-0.556	274.35	0.674	4	0.954518
10	0.647	0.115	-2.758	274.35	10.088	4	0.03897

Legend: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies evidence based practice 20 = assesses risk; SE.= standard error; FitResid = Fit residual; DF= degrees of freedom; ChiSq= Chi Square; Prob= probability.

4.3.10 Person separation index (PSI)

The PSI was 0.93 indicating the ability to discriminate between four or more levels of performance.

4.3.11 Dimensionality

Analysis of the pattern of item loadings on the first extracted factor of the residuals shows that the residuals loaded in opposite directions on two subsets defined by positive and negative loadings on the first factor. Only those items with loadings greater or less than 0.3

were considered. The next step was to investigate if the person estimates (location values) based on scores that underpin each of these sets of items were significantly different using independent t-tests. A confidence interval for a binomial test of proportions was calculated for the observed number of significant tests. In the pilot trial data 23 cases out of 293 (7.84% or .07 (95% CI 0.05 - .1)) had statistically different scores on each of the subsets of items. As the range of probabilities estimated for the 95% confidence interval around the observed .07 contains the value 0.05, unidimensionality of the scale is supported (Smith, 2002).

4.4 Discussion

Application of the Rasch measurement model in this pilot trial provided preliminary support for the APP instrument as a measure of professional competence of physiotherapy students in the work place. In particular, evidence from Rasch analysis justified the continuation of research into the field testing phase.

On each item, students were rated on a five level response scale from poor to excellent demonstration of competence. The expectation was that as student ability increased the probability that they would be rated at a higher level would increase in an ordered fashion from low to high performance. Analysis of the APP pilot trial data demonstrated that educators were using the five level response scale as intended. The scale exhibited high reliability (PSI=.931) with no disordered thresholds (Figure 4.2). A PSI of 0.9 indicates a highly discriminative scale able to detect four or more levels of student ability.

Overall fit to the Rasch model was adequate with no individual item or person misfit of major concern. Investigation of person fit residuals identified 3 of the 295 completed assessments (1%) with unexpected scoring on items 18 and 19 (EBP and discharge planning). Although this indicates a few exceptions to good fit, seeking feedback from the clinical educators regarding these items prior to field testing would provide insight into difficulties the educators might have experienced in interpreting and scoring these items.

The person-item threshold distribution graph in Figure 4.3 illustrates the items were sufficiently targeting the intended performance. However, at the right hand end of the graph there are several item thresholds that even the most able students could not achieve and at the left hand end there are several item thresholds that were too easy for the least able students. This result may be due in part to the minimal level of training in completion of the instrument provided to the clinical educators prior to the start of the pilot trial. This aspect of targeting needs to be more fully investigated during field testing when educator training will be comprehensive and standardised.

Table 4.3 shows the sequence or hierarchy of average difficulty of the 20 competencies on the APP. This provides an indication of which clinical competencies may be easier to acquire e.g. communication and professional behaviours and those that are more difficult and therefore may be expected to take longer to master e.g. application of evidence based practice, analysis and planning (critical thinking).

Previous research has demonstrated that other physiotherapy work-based assessment instruments appear to tap more than one underlying construct. The internal structure of 3 instruments has been investigated using principal components analysis (PCA). The Clinical Performance Instrument (CPI) (Task Force for the Development of Student Clinical Performance Instruments, 2002), PT MACS (Logemann, 2006; Stickley, 2005) and the Clinical Internship Evaluation Tool (CIET) (Fitzgerald, et al., 2007) appear to be multidimensional instruments, measuring at least two constructs, physiotherapy specific clinical skills and professional behaviour. The APP includes items relating to communication, professional behaviour and physiotherapy specific skills but nevertheless appears robust when tested against the assumptions of the Rasch measurement model, with the summary fit statistics and independent t-test analysis supporting the assumption of unidimensionality. These results are similar to those of (Conaghan, Emerton, & Tennant, 2007). They used Rasch analysis to investigate the Oxford Knee Scale and demonstrated that, despite the scale measuring the two attributes of pain and function, it was shown to be unidimensional and hence was potentially measuring a construct that was reflected in both attributes.

There are a number of limitations to this study including the absence of differential item functioning analysis, low levels of training provided to the assessors, and the lack of

qualitative data on the utility and acceptability of the instrument to both educators and students.

When developing an instrument for assessment of work-based performance, the research situation demands participation and collaboration. Checkland and Holwell (1998) argued that action research best incorporates these factors. It is a method for yielding simultaneous action (change) while conducting quality research. The results of this type of research are practical, relevant, and improve professional practice through continual learning and progressive problem solving. The cyclical research paradigm of participatory action research is an essential aspect of best practice in instrument development. However, this approach also imposes a significant burden on stakeholders involved in the repeated research cycles of pilot and field testing. A pragmatic decision was made not to impose unnecessary work load burdens on the clinical educators during the pilot trial and to clarify functioning of the items, scale and overall instrument dimensionality before undertaking additional investigation.

Wilson (2005) recommended conducting think aloud and exit interviews if appropriate levels of validity evidence for the instrument are to be compiled. These interviews enable investigation into whether educators were interpreting and using the items, performance indicators and response scale as intended. They were not conducted during the pilot trial but were planned for the field testing phase.

It has been argued that the Rasch Measurement Model is the standard for psychometric evaluations of outcome scales, and should be used during the development phase or when reviewing the psychometric properties of existing instruments (J. Hobart & Cano, 2009; Tennant & Conaghan, 2007). A systematic review of instruments to assess professional competence in physiotherapy students (Chapter One) revealed that only three of the eight instruments located had been investigated using IRT or Rasch analysis. The CPI and PT MACS were investigated by Logemann (2006) using a 2-parameter graded response IRT model and the Student Clinical Competence Scale using Rasch analysis by Rheault and Coulson (1991). The reported IRT and Rasch analyses of these three instruments were not comprehensive enough to enable an informed comparison to results of this study.

4.5 Chapter Summary

This chapter has presented an analysis of the pilot data, which has indicated that the APP data had adequate fit to the chosen measurement model (Rasch Partial Credit Model), the rating scale was operating as intended, the items were sufficiently targeting the intended performance and the instrument could discriminate at least four levels of competence. The results of the pilot testing allowed the first field test of the APP to proceed with confidence. Chapter Five provides the qualitative results following the pilot trial of the newly developed instrument prior to the first field test, aiding the continued development of the APP for use as a national assessment instrument.

5. Chapter Five: APP Pilot Trial - Qualitative evaluation

5.1 Introduction

Rasch analysis of the pilot data presented in Chapter Four indicated that APP data had adequate fit to the Rasch partial credit model, the Person Separation Index demonstrated the scale was internally consistent discriminating between four groups of students with different levels of professional competence, the items were targeting the intended construct (professional competence) and the instrument demonstrated unidimensionality.

Feedback from stakeholders provides important information about the instrument that is not provided by quantitative data analysis, such as insight into item interpretation, the acceptability/utility of the instrument to the end users, educational impact and unintended consequences of instrument use. Stakeholder feedback also provides information on training requirements enabling further development and refinement of appropriate training resources prior to field testing. Feedback from clinical educators who used the APP instrument during the pilot trial and from other members of the profession was collected through focus groups and presentations conducted after pilot testing and prior to commencement of field testing. This chapter presents the analysis of qualitative data gathered during the pilot trial.

5.2 Method

As discussed in Chapter Three (section 3.3), the research group considered stakeholder support to be a critical factor in the development of an instrument that would be used by practitioners to assess competence to practice. This research needed to engage stakeholders from around Australia to maximise the sense of ownership by the profession in the final product. Focus groups, one to one interviews, surveys, workshops, email, and teleconferencing were scheduled across the duration of the research to gather the input of as many stakeholders as possible. This approach enabled triangulation and reinforcement of decisions based on qualitative data obtained from multiple sources.

5.2.1 Focus groups

Focus groups were conducted following the pilot trial to gather feedback on the APP version 2, to develop the clinical educator and student surveys to be used during field testing, and to ascertain training requirements for educators and students prior to field testing. The methods used to conduct the focus groups were the same as those described in Chapter Three, section 3.3.1 and are not repeated here except where the procedures varied.

5.2.1.1 Focus group questions

In addition to the questions used in earlier focus groups (section 3.3.1.6), there was modification and refinement of questions from one focus group to the next (Cote-Arsenault & Morrison-Breedy, 1999). Hence, questions for each focus group evolved across the groups with results from the first three groups giving rise to topics that warranted discussion in subsequent groups. Questions were modified or extended particularly where saturation of answers / information was achieved.

5.2.2 Presentations

In addition to the structured focus groups, four presentations on the APP were conducted. The sessions took the format of a PowerPoint presentation outlining progress in the research and presenting the Rasch analysis from the pilot trial. The presentation was followed by questions and discussion.

The aims of the presentations and discussions were to:

- disseminate information about the project and inform stakeholders
- gather opinions and feedback from stakeholders on the instrument and its use
- engage the physiotherapy profession in the research.

The presentation sessions were not recorded. Participants at the presentations were informed about the research and consent for a research assistant to take notes during the discussion were sought. A copy of the discussion summary was available on request

following each presentation. All participants consented and no request for summary notes was received.

Following completion of all qualitative data collection, a summary document was disseminated to participants and more broadly to the wider profession via email, an Australian Physiotherapy Association (APA) e-newsletter and an article in the APA publication, *InMotion* (Appendix 5.1). The document described progress in the research, pilot trial results and modifications made to the APP instrument based on the pilot trial data. Circulation of the document disseminated information about the research and maintained engagement of the profession in the research.

5.2.3 Data analysis

In place of the manual qualitative data analysis conducted previously (section 3.3.1.7), a content analysis of the transcripts and field notes was conducted using The Leximancer 2.25² (2009) software. The analysis was conducted in two stages. The first was an analysis of the manifest content (quantitative assessment). This procedure totals the frequency of issues discussed (removing potential bias of the facilitators in otherwise determining the content). The second was an analysis of the latent content (qualitative assessment). Leximancer software pinpoints evidence in the transcript (via linkages and co-occurrences of concepts) to support conclusions drawn in regard to changes, improvements and strengths of the developing instrument. Leximancer 2.25 (2009) also allows the researchers to systematically link concepts. For example, by clicking on the word 'student', the co-occurrences of concepts located in the text in close proximity to 'student' are mapped. While analysis of the manifest content describes the most frequently raised issues, analysis of the latent content supports decisions regarding why these issues have become important and what actions are appropriate. The data was colour coded into the issue categories identified by the researchers prior to the group. Figure 5.1 illustrates the different stages through which Leximancer 2.25 (2008) software processes data that eventually leads to the concept map.

² Leximancer 2.25, Leximancer Pty Ltd, , www.leximancer.com, 2008.

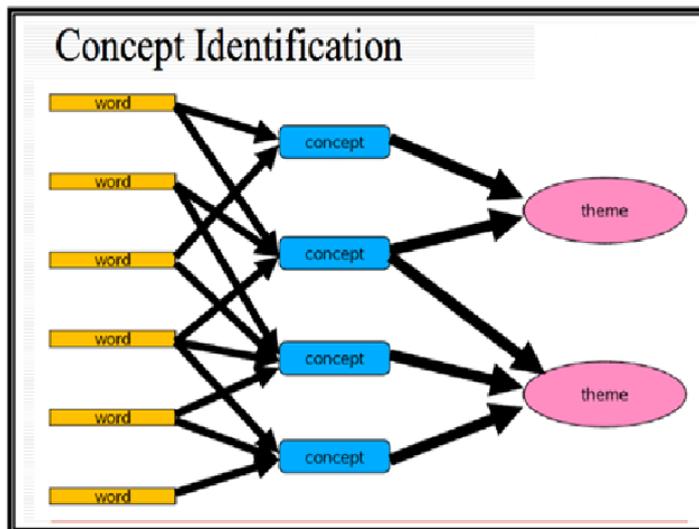


Figure 5.1: Concept map development in Leximancer (Kivunja 2009).

5.3 Results

5.3.1. Participants

Ten focus groups (Table 5.1) with a total of 104 participants were conducted, in addition to four presentations/discussions with a further 67 members of the profession (Table 5.2).

Each focus group included participants who represented a broad cross section of clinical areas of practice. Participants confirmed that the transcripts accurately reflected the focus group discussion and no requests for amendments to the transcript were made.

Qualitative data from focus groups and presentations were analysed individually and then collated and sorted by topic and content.

Table 5.1: Focus group participants (after pilot trial and prior to Field Test One)

Focus Group No./ Date	Setting/ Facility type	No of participants /gender	Job description of participants: Clinical Educators (CEs), Academic, Manager	Years of experience as a CE mean years, (SD)
1 Dec 2006	QEI Hospital, Brisbane, QLD Metropolitan Public Hospital	13 (7F, 6M)	CEs	9.7 (8.0)
2 May 2007	Ballina Base Hospital, Ballina, NSW Regional Public Hospital and Community Health Centre	9 (8F, 1M)	Managers of Northern Rivers Health District Hospitals, NSW	20.2 (9.7)
3 May 2007	Westmead Hospital, Sydney, NSW Metropolitan Public Hospital	12 (7F, 5M)	CEs and managers	10.2 (6.4)
4 May 2007	Prince of Wales Hospital, Sydney, NSW.	11 (8F, 3M)	CEs and managers	9.6 (7.1)
5 June 2007	Metropolitan Public Hospital Townsville Hospital, Townsville, QLD	6 (4F, 2M)	CEs and managers	6.4 (2.3)
6 June 2007	Regional Public Hospital James Cook University, Townsville, QLD	4 (4F)	Academics	13.2 (7.4)
7 April 2007	Dunedin University, Dunedin, NZ Australia and New Zealand Clinical education managers meeting	14 (10F, 4M)	Academics	18.4 (11.2)
8 Feb 2007	Griffith University, Gold Coast, Qld	11 (8F, 3M)	CEs, private practice and hospital based	5.1 (2.8)
9 Feb 2007	La Trobe University, Melbourne, Victoria	11 (7F 2M)	CEs and managers	13.2 (8.3)
10 March 2007	Box Hill Hospital, Box Hill, Victoria. Regional Public Hospital	13 (8F 5M)	CEs	10.3 (7.6)

Code: CE=clinical educator, NZ= New Zealand, SD= standard deviation, m=male, f=female

Table 5.2: Presentations on APP results (after pilot trial before field test 1)

Facility and date	No. of participants	Professional duties
Toowoomba Hospital, Toowoomba, QLD June 2007	11	Clinical educators
Australia and New Zealand Clinical Education Managers (Reference Group) and Heads of Schools meeting Dunedin, April 2007	17	Academics and health service managers
Griffith University School of Physiotherapy and Exercise Science staff of school of physiotherapy Gold Coast April 2007	14	Academics and clinical educators
The University of Queensland School of Health and Rehabilitation Sciences staff seminar presentation, St Lucia, May 2007	25	Academics and clinical educators

5.3.2 Analysis of the manifest content (quantitative assessment)

The Leximancer 2.25³ automatically identifies key themes, concepts and ideas by data mining large amounts of text, and visually represents information in ‘concept maps’ showing the main relationships. These relationships can be examined in more detail by exploring major connections. The first analysis of the focus group transcripts in Leximancer generated the concept map shown in Figure 5.2

³ Leximancer 2.25, Leximancer Pty Ltd, , www.leximancer.com., 2008.

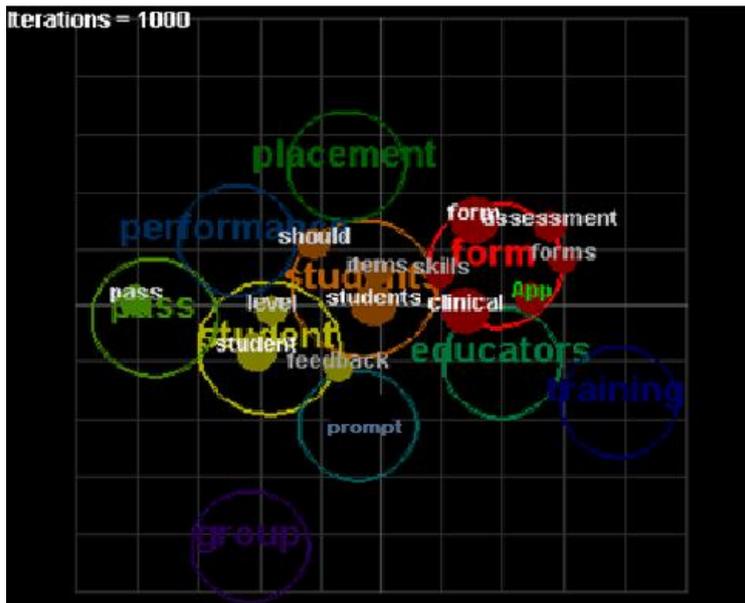


Figure 5.2: Concept map displaying frequency of concepts arising from focus group transcripts

The frequency of each concept is represented by the boldness of the text. Intersecting circles represent an overlap between themes, while the relationship between concepts is demonstrated by the size of concept point. Circles that are further apart (eg training, group) show themes that appear independently in the data. In addition to themes and concepts, the first stage of analysis also generates statistical data which can be used to analyse the relative and absolute frequencies of all the concepts (Table 5.3).

Table 5.3: Initial concept ranking from focus groups

Concept	Absolute Count	Relative Count
<u>student</u>	85	100%
<u>form</u>	69	81.1%
<u>clinical</u>	62	72.9%
<u>students</u>	55	64.7%
<u>pass</u>	43	50.5%
<u>assessment</u>	43	50.5%
<u>should</u>	41	48.2%
<u>level</u>	39	45.8%
<u>APP</u>	38	44.7%
<u>placement</u>	30	35.2%
<u>forms</u>	27	31.7%
<u>items</u>	26	30.5%
<u>skills</u>	25	29.4%
<u>feedback</u>	25	29.4%
<u>scale</u>	24	28.2%
<u>prompt</u>	23	27%
<u>fail</u>	23	27%
<u>grade</u>	22	25.8%
<u>unit</u>	20	23.5%
<u>educators</u>	20	23.5%
<u>performance</u>	19	22.3%
<u>training</u>	18	21.1%
<u>analysis</u>	18	21.1%
<u>instrument</u>	16	18.8%
<u>testing</u>	15	17.6%
<u>group</u>	15	17.6%

Legend: APP= Assessment of Physiotherapy Practice instrument

Concept editing was used to merge duplicates and remove irrelevant concepts, such as, form/forms (merged), should (removed), student/students (merged), assessment/testing (merged).

The concept map (Figure 5.2) and the concept-ranking table (Table 5.3) identified the most frequently raised concepts in descending order. The second most frequently raised concept was identified as the word 'form' which was also represented by the words 'instrument' and 'tool'. These words combined represent the most frequently raised concept, followed by student with training being among the least frequently mentioned.

5.3.3 Analysis of the latent content (qualitative assessment)

While these frequency totals go some way to confirm what the researchers identified as the most commonly raised issues of the focus groups, greater depth of understanding was required to inform the developing instrument. Therefore, the Leximancer software was used to pinpoint evidence in the transcripts to investigate what the participants perceived to be required in regard to changes, improvements and strengths of the developing instrument. Leximancer allowed the researchers to systematically link concepts. For example, by clicking on the word 'student', the co-occurrences of concepts located in the text in close proximity to 'student' are mapped and are shown in Figure 5.3.

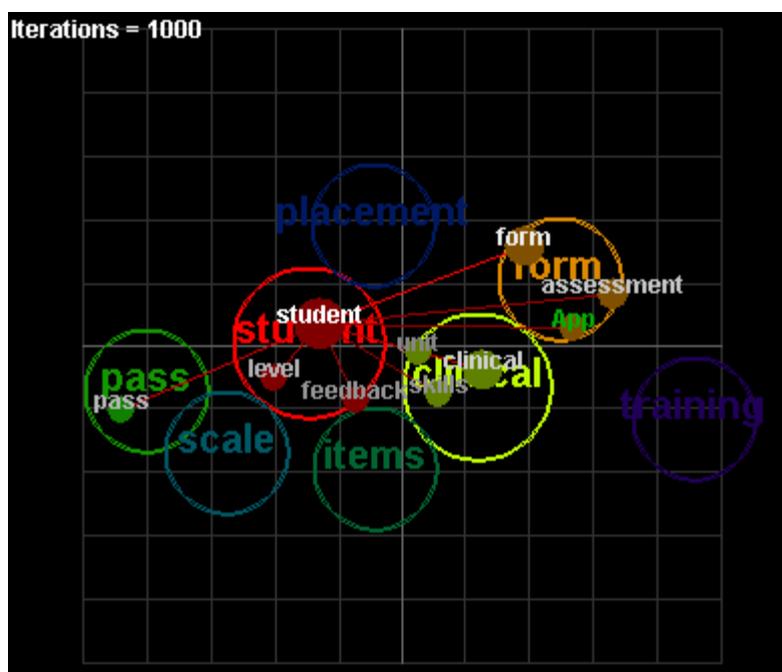


Figure 5.3: Concept map showing co-occurrences of the word 'student'

Table 5.4 shows the order and frequency with which the other major concepts appear in the text in close proximity to the concept 'student'. For example, the concepts 'student' and 'form' co-occur in the text 45 times. Also, the concept 'pass' appears in the text in close proximity to the concept 'student' a total of 31 times reflecting the level of interest in the passing level to be used for rating items on the form (instrument).

Table 5.4: Initial concept ranking of co-occurrences of the word 'student'

Concept	Absolute Count	Relative Count
<u>form</u>	45	34.6%
<u>pass</u>	31	23.8%
<u>clinical</u>	28	21.5%
<u>level</u>	27	20.7%
<u>unit</u>	19	14.6%
<u>assessment</u>	18	13.8%
<u>feedback</u>	17	13%
<u>grade</u>	16	12.3%
<u>APP</u>	15	11.5%
<u>placement</u>	15	11.5%
<u>scale</u>	15	11.5%
<u>skills</u>	14	10.7%
<u>items</u>	14	10.7%
<u>fail</u>	13	10%
<u>performance</u>	11	8.4%
<u>prompt</u>	9	6.9%
<u>analysis</u>	8	6.1%
<u>educators</u>	8	6.1%
<u>training</u>	3	2.3%

This feature in Leximancer allowed the researchers to identify related concepts and then locate them in the text to explore the meaning and context of this co-occurrence. For example, linking the concepts 'student' with 'grade' generated the following evidence from the text:

Responsibility of grading:

"It is good that as educators we do not have to give the students a final mark using this form. However, the students can easily figure it out. If each criterion were weighted differently though, it would be more difficult for the students to figure out their mark."

"It is preferable that the educator only needs to inform the student if they passed or failed and the university then decides the grade. This removes the pressure from the educator (particularly younger educators). Placement 'is about learning, not grading'".

Linking the concepts of 'scale' and 'pass' reveals the following text from two focus group participants.

Scale size

"5 levels on the scale are good but having only one for the pass level is not ideal. This allows for 2 x failing, a pass, and 2 x good grades. Perhaps another level on the scale would be helpful".

"It is recognised that most educators can differentiate between 4 - 5 levels of ability. Any greater makes the scale unnecessarily complex and does not improve the instrument".

Scale wording and prompting

"The scale simplifies things for clinicians but use of the word prompting will cause confusion".

"What is prompting and what is collaborative discussion?"

A systematic approach was taken by the researchers to capture co-occurrences of all concepts identified by Leximancer as frequently appearing in the text. As an example, beginning with the concept 'student' (the most frequently appearing concept) the researchers captured all co-occurrences of 'student' with the remaining most frequently appearing concepts in rank order by frequency (form, pass, clinical, level, unit etc). Then the researchers captured the co-occurrences of the second most frequently appearing concept 'form' with the remaining most frequently appearing concepts (pass, clinical, level, unit etc) and so forth until all concept co-occurrences had been captured. Although presented as a linear, step-by-step procedure, the analysis was an iterative and reflexive process. This process enabled the researchers to confidently establish frequently raised issues (manifest content analysis). By linking the concepts to locate and retrieve their co-occurrences from the text, deeper contextual understanding of why these concepts/issues were frequently raised was obtained (latent concept analysis). The evidence was then colour-coded and categorised and considered in the context of issues relevant to instrument development and the questions asked at each focus group. The process also aided identification of consensus and disagreement in each focus group.

A summary of the target issues, actions arising and decisions on completion of the ten focus groups and four presentation/discussion sessions are presented in Table 5.5.

5.3.4 Definition of entry level / passing standard performance

As the focus groups progressed, participants requested a definition of entry level/passing standard performance. To address this, participants in the last three focus groups were asked to complete this statement: "A student is performing at the entry level standard when they....". Four examples of the definitions provided by participants are provided below.

1. *"A student is performing at the entry level standard when they can conduct an effective assessment and treatment, plan management for a patient independently (seeking guidance as necessary), with professional behaviour throughout, in a safe manner."*
2. *"A student is performing at the entry-level standard when they can demonstrate the ability to conduct an assessment that includes: history taking, functional assessment, impairment assessment. They can then use this information to prioritise the patient's main problems and devise and implement appropriate safe treatments."*
3. *"An entry level physiotherapist is one who is competent according to the Australian Standards i.e. one who is competent to manage a reasonable workload and to assess, interpret, implement and evaluate basic treatment in a range of therapeutic areas. This means that they are at a level that is competent but yet may still require supervision."*
4. *"A student is performing at the entry level standard when they demonstrate a resourceful/self-directed approach to continued learning. A professional awareness and commitment to effective, efficient, safe and competent assessment and management of patient's problems. They will have the attributes of a reliable/responsible team-player and know their limitations."*

All definitions provided were transcribed, coded and analysed by the research group for key words and themes. The definition of entry level performance agreed to was:

“A student is performing at the entry level standard when they demonstrate an understanding of patient centered physiotherapy practice and are able to manage a variety of patients such that the major problems are identified, goals established and intervention is completed safely, professionally, effectively and within in a reasonable time frame. While achieving this, the student is aware of their limitations and where to seek assistance.”

This definition was subsequently included in all training resources provided to participants in the first field test. In line with the action research approach, the definition would be reviewed by focus groups conducted following the first field test.

5.3.5 Addition of global rating scale (GRS)

In the initial version of the assessment instrument, individual item based (analytic) scoring was chosen as this was similar to current practice within the physiotherapy profession. A combined or hybrid approach utilising both scores for specific items and a global rating scale (holistic scoring) has been proposed to be the most suitable method of evaluating undergraduate clinical competence of medical students (Ringsted, et al., 2003). The global rating provides additional information regarding overall perception of ability that may not be captured by item scores (Friedman Ben-David, 2005; McIlroy, et al., 2002; P. J. Morgan, et al., 2001). The addition of a global rating scale was discussed in the final three focus groups. As participants had no previous experience using global rating scales to assess student performance, participants were undecided regarding the potential value of adding a global rating scale. Based on published arguments a global rating of performance scale was added to the requirement to rate individual items. Global perceptions of typical performance provide insight into the validity of tallying item scores to provide an overall grade for performance, as some alignment between total item scores and global rating is anticipated. Global rating of performance is a reliable measure of clinical skill that is also sensitive to increasing levels of expertise and is considered a good measure of complex interactions such as communication, rapport and other professional behaviours (Daelmans, et al., 2005; Domingues, Amaral, & Zeferino, 2009).

5.3.6 APP (version 3) for use in Field Test One

Following consideration of the pilot trial quantitative and qualitative data, a number of changes were made to the APP (version 2). A summary of the modifications is presented in Table 5.6. The amended APP instrument (version 3) for use in Field Test One is provided in Appendix 5.2.

Table 5.5: Summary of focus group data following pilot trial

Target issue and discussion	Outcomes/ Decisions/ Actions arising	Group decision
<p>Training</p> <ul style="list-style-type: none"> • <i>Timing of training workshops</i>: best conducted just prior to the clinical education placements. This helps to iron out via discussion any difficulties being experienced and streamline the training package. • <i>Standardisation</i>: training will help standardise use of the instrument. • <i>Expertise in grading</i>: CEs should not assign the final grade to the student. This should be done by the university staff. • <i>Training format</i>: requests for training package educators could complete on their own if they are unable to attend a workshop. Request also for summary of main issues associated with completing the instrument so if pressed for time, at least CEs would read these pages. Suggested this be in a frequently asked questions (FAQs) format. • <i>Information in training package</i>: guidance on how to give feedback, how to use the scoring system, definition of a pass standard performance. 	<ul style="list-style-type: none"> • Development of training package/resource manual to meet clinical educator needs (comprehensive, practical and accessible). • Training to coincide with scheduled clinical educator meetings wherever possible. • Inter-rater reliability trials will be conducted • Clinical educators are not required to assign the final grade, this will be done by University academics. • A FAQs section incorporated into the training package 	<p>Consensus on outcomes</p>
<p>Items</p> <ul style="list-style-type: none"> • Difficulty assessing items on discharge planning and evidence based practice experienced by some educators in pilot trial 	<ul style="list-style-type: none"> • More information on these items and how to assess them to be included in training resource. 	<p>Majority</p>
<p>Scale size</p> <ul style="list-style-type: none"> • <i>Number of scoring categories</i>: discussion concerning increasing number of scoring categories or maintaining 5 categories. • <i>Pass/fail categories</i>: discussion regarding whether some items could be graded pass/fail only, eg., risk management. • <i>Student acceptance of scale size</i>: will students want more scoring categories to highlight change in performance? • Instead of not scoring an item, category for not assessed requested 	<ul style="list-style-type: none"> • Rasch analysis of pilot trial results showed scale was working appropriately, therefore agreement that 5 level scale to remain with 2 failing levels and 2 passing levels. Review following field tests. • Consideration of having both a 5 level scale for some items and a pass/fail scale for some items. Review following field test. • Not assessed (n/a) category added in 	<p>No consensus on scale size but agreement to review after field tests</p>
<p>Scale wording</p> <ul style="list-style-type: none"> • The term ‘some’ in the description of a rating of 1 thought to be too positive. • Terms prompting, complex and non complex patient conditions creates confusion for assessors 	<ul style="list-style-type: none"> • Description of rating 1 changed to read: ‘demonstrates few performance indicators to an adequate standard’. • Reference to prompting and complex/non-complex removed. Review after field test. 	<p>Consensus</p>

Target issue and discussion	Outcomes/ Decisions/ Actions arising	Group decision
<p>Performance indicators</p> <ul style="list-style-type: none"> • <i>Demonstration of performance indicators</i>: discussion regarding use of performance indicators as a checklist or as a guide to expected behaviours • <i>Use of performance indicators in assessment</i>: very positive feedback that PIs are helpful to guide mid unit feedback and assist end of unit summative grading. • <i>Training in use of PIs</i>: correct use of PIs needs to be included in training 	<ul style="list-style-type: none"> • Training to stipulate that the performance indicators are not an exhaustive list or checklist but rather a guide to the kinds of behaviours the clinical educator would look for when scoring a particular item. • Training to stipulate that PIs may be used to inform the clinical educator of the kinds of behaviours students should observe in clinical educators during the clinical placement 	Consensus
The use of 0 on the scale	<ul style="list-style-type: none"> • Training to include information on correct use of PIs • Retain '0' at this stage because it allows for very poor grades and is also useful for mid-unit feedback. 	Review after field test
Mid-unit uses e.g. feedback	<ul style="list-style-type: none"> • Training manual instructions to include that APP is used for formative feedback at mid unit and summative assessment at end of unit. The APP and PIs provide a good framework for formative feedback at mid-unit. 	consensus
<p>The use of entry-level standard for passing and defining entry-level performance Video examples of entry level performance requested</p>	<ul style="list-style-type: none"> • Definition of entry level performance i.e. average rating of 2(pass standard) included with instrument and in training materials. Definition described by the focus group participants (refer to section 5.3.3). • Students must achieve entry-level standard to pass each clinical unit 	Consensus
<p>Should students have to pass all sections to pass overall:</p> <ul style="list-style-type: none"> • For example, should risk management have to be passed for student to pass overall? • Passing professional behaviour and communication sections and failing the physiotherapy skills sections. Would an overall pass in this case be acceptable? 	<ul style="list-style-type: none"> • Further testing and analysis will review the relationships between section scores 	Review following field testing
Applicability of APP across placement settings and clinical areas	<ul style="list-style-type: none"> • The APP will be tested for bias including variables such as placement setting, gender, experience level of clinical educator using differential fit analysis option in RUMM software • The language will be reviewed to ensure suitability for use in all placement settings. 	Consensus

Target issue and discussion	Outcomes/ Decisions/ Actions arising	Group decision
<p>Other issues</p> <ul style="list-style-type: none"> • Current situation is that there is pressure on clinical educators because of university specific clinical assessment forms in use. Those that are time consuming result in clinical educators reducing the number of students they can manage/supervise. • Addition of a global rating of performance scale 	<ul style="list-style-type: none"> • Participants agreed that a single instrument used by all universities would ease educator burden and improve likelihood of being able to increase student placements. • Participant experience with scales was limited and few opinions offered on the merit of adding a global rating scale 	<p>Consensus</p> <p>GRS added</p>

Table 5.6: Modifications to APP (version 2) following pilot trial

Requested modifications to APP (v1)	APP (version 2) used in pilot trial	APP (version 3) used in field test 1
Change wording on scoring scale descriptors	<p>0 = Infrequently demonstrates performance indicators, requires constant prompting with usual/typical (non-complex) patient presentations</p> <p>1 = Demonstrates some performance indicators to an adequate standard, requires frequent prompting with usual/typical (non-complex) patient presentations</p> <p>2 = Demonstrates most performance indicators to an adequate standard, requires prompting in atypical or complex patient presentations</p> <p>3 = Demonstrates most performance indicators to a high standard, requires occasional prompting</p> <p>4 = Demonstrates most performance indicators to an excellent standard, rarely requires prompting</p>	<p><i>0 = Infrequently/rarely demonstrates performance indicators</i></p> <p><i>1 = Demonstrates few performance indicators to an adequate standard</i></p> <p><i>2 = Demonstrates most performance indicators to an adequate standard</i></p> <p><i>3 = Demonstrates most performance indicators to a good standard</i></p> <p><i>4 = Demonstrates most performance indicators to an excellent standard</i></p> <p><i>n/a = not assessed</i></p>
<p>Relocate scale definitions to top of the instrument</p> <p>Add in scoring rules to instrument to guide assessors</p>	<p>Scale definitions at bottom of APP (v2) instrument</p> <p>No scoring rules on instrument</p>	<p>Scale definitions at top of APP (v3) instrument</p> <p><i>Scoring rules added:</i></p> <ol style="list-style-type: none"> 1. Circle n/a (not assessed) only if the student has not had an opportunity to demonstrate the behaviour 2. If an item is not assessed it is not scored and the total APP score is adjusted for the missed item. 3. Circle one only number for each item 4. If a score falls between numbers on the scale the higher number will be used to calculate a total. 5. Evaluate the student’s performance against the competency level expected for a beginning physiotherapist <p><i>Item 7: Conducts an appropriate patient/client interview</i></p> <p><i>Item 9: Performs appropriate physical assessment procedures</i></p>
Modify wording on items 7 and 9	<p>Item 7: Conducts an appropriate patient/client interview (subjective assessment)</p> <p>Item 9: Performs appropriate assessment procedures (objective assessment)</p>	

Add in global rating scale	No global rating scale, individual items rated	<i>Global rating scale added: In your opinion as a clinical educator, the overall performance of this student in the clinical unit was: Poor; Adequate; Good; Excellent</i> APP (version 3) used in field test 1
Requested modifications to APP (v1) Modify wording on PIs to ensure comprehensive, clear, and appropriate for all clinical areas.	APP (version 2) used in pilot trial 1. Demonstrates an understanding of patient/client rights and consent • <i>manages time and resources effectively</i> 3. Demonstrates practice that is ethical and in accordance with relevant legal and regulatory requirements • <i>Treats patients/clients within scope of expertise</i> 5. Communicates effectively and appropriately – verbal/non-verbal • <i>Uses suitable non-medical terminology and avoids jargon</i> 9. Performs appropriate assessment procedures (<i>objective assessment</i>) 17 Progresses intervention appropriately • <i>Implements safe and sensible treatment progressions</i>	1. Demonstrates an understanding of patient/client rights and consent • <i>Manages time and resources effectively (removed as is not a relevant PI for this item)</i> 3. Demonstrates practice that is ethical and in accordance with relevant legal and regulatory requirements • <i>Understands scope of expertise</i> 5. Communicates effectively and appropriately – verbal/non-verbal • <i>Uses suitable language and avoids jargon</i> 9. Performs appropriate assessment procedures (physical assessment) 17 Progresses intervention appropriately • <i>Demonstrates or describes safe and sensible treatment progressions</i>
Format PIs onto 2 pages	PIs on 5 pages	<i>PIs formatted onto 2 pages</i>

(Cod

e: amendments to instrument highlighted in *italics*)

5.4 Discussion

Overall, focus group participants demonstrated strong convergence in their opinions regarding assessment design and training regardless of whether they were clinical educators, university academics or managers. Consensus on common issues may have been facilitated by homogeneity of participants within and between groups.

Participants agreed the addition of two items provided more comprehensive coverage of the professional competency content that required assessment during workplace placements. Similarly consensus was reached on the rewording of the scoring scale. Participants recommended removal of references to the degree of prompting considered necessary by the educator and the complexity of the patient's condition to improve rating scale clarity and reduce confusion for assessors. As in earlier focus groups, participants discussed the number of categories on the scoring scale. Rasch analysis of pilot trial data had demonstrated that assessors were using the five point scale appropriately to differentiate increasing levels of performance. In view of these results, participants agreed to retain the five level scale and entry level performance as the passing standard and review again following field testing.

Consensus was also reached on the addition of a 'not assessed' category for each item. Some participants were concerned that the not assessed category could be used by an educator as a default rating for an item where a student has not had sufficient opportunity to achieve an entry level standard of performance on that item. Participants agreed appropriate training in use of the instrument would address this issue. Several participants identified confusion relating to how item 19 (applies evidence based practice in patient/client care) was to be assessed. This finding reflects those of earlier studies by Cross et al (2001) where behaviours such as 'demonstrating research knowledge' proved to be difficult for educators to observe and assess. Cross and colleagues however cautioned against removing items such as this from an assessment instrument as it could "*waste an opportunity to foster an evidence-based culture among particular groups of practitioners*" (p 349). If assessment truly drives learning, this appears to be the preferred approach.

More information on this item and how to assess it would be included in future training resources.

The performance indicators were considered extremely helpful when providing formative feedback to students and developing strategies to improve performance. Participants commented the indicators provided greater transparency to students regarding the expected standard of performance and reduced the tendency of assessors to rely on personal interpretation of the physiotherapy standards when assessing clinical practice. The criteria for the performance indicators recommended by Cox (1996) and Streiner and Norman (2003) that they must describe expected behaviours, need to be unambiguous, observable, and measurable appeared to have made them useful.

Focus group participants supported the concept that any instrument used to assess professional competence should function consistently irrespective of the experience of the clinical educators, gender of the students or clinical area of the placement. Investigation of differential functioning of the items requires collection of detailed demographic data from both educators and students. Focus group participants agreed this was an essential component of the next phase of research.

Questions and targets for discussion for each focus group evolved across the ten groups. Topics generated in the first three focus groups were pursued with groups that followed. One topic was the definition of entry level performance standard. The definitions proposed by the participants of the final three groups demonstrated a high level of agreement on behaviours considered representative of entry level performance. As the focus group questions evolved so did the priority conferred on each issue discussed. Standardisation of training in assessment practices and use of the APP were more frequently discussed in later focus groups when participant interest in the field test phase of the research had increased.

All participants agreed on the importance of training clinical educators and students in use of the assessment instrument. Themes relating to preparation of training material included accessibility, standardisation, time efficiency, variety in modes of training delivery, clarity and succinctness of information, definitions of performance standards, role of formative and summative assessment, involvement of students in self assessment, exemplars of

performance standards and practise in use of the instrument. These issues will be incorporated into training resources to be developed prior to field testing.

A systematic review of the literature (Chapter One) identified that current practice in assessment of clinical performance of physiotherapy students in Australia and New Zealand involved the use of individual item (analytic) scoring only. This was the scoring system initially chosen for the first version of the assessment instrument (CAPS v1). Global (holistic) scoring is suited to measurement of process aspects of a task also described as higher components of clinical competence, for example, rapport building, critical thinking and risk minimisation (Hodges, McNaughton, Regehr, Tiberius, & Hanson, 2002; McIlroy, Hodges, McNaughton, & Regehr, 2002; Regehr, Freeman, Robb, Missiha, & Heisey, 1999). As such, the student is measured on components of the task that relate more closely to the quality of the performance of the task, such as the ability to generate rapport, the general approach to the task, professional behaviour and patient education skills, rather than the actual practical demonstration of individual skills.

Domingues et al.(2009), Morgan et al (2001) and McIlroy et al (2002) demonstrated that assessors assigned higher ratings on the humanistic domains compared with technical skill domains when using a global rating scale. Domingues et al.(2009) suggest a combined or hybrid approach utilising both specific items and global rating scales may be the most suitable method of evaluating clinical competence through the ability to adopt the strengths and address the limitations intrinsic to each of the rating scales when used in isolation. This research and the ability to examine the relationship between item scoring and rating of overall performance led the research group to include a global rating scale on the APP instrument for investigation during field testing.

Following the focus groups and presentations, a summary of the modifications made to the APP (version 2) instrument was circulated to all participants and across the profession through physiotherapy publications. Dissemination of information provided participants with evidence of their input into development of the instrument and related training resources and helped maximise the sense of ownership of the profession in the final product. The following Chapter presents the first phase of field testing.

6. Chapter Six: Field Test One - Qualitative evaluation

6.1 Introduction

The primary aim of this research was development of an instrument to assess professional competence of physiotherapy students in the work place. To achieve this aim and to ensure adequate evidence of validity, collection of qualitative and quantitative data within a cyclical research paradigm of planning, acting, monitoring and evaluating was implemented.

Phase One of the research (Chapters 2 and 3) described the development of the assessment instrument and Phase Two (Chapters 4 and 5) presented the results of the pilot trial of the newly developed instrument. The pilot trial version of the assessment instrument consisted of 20 items assessed by clinical educators using a five point rating scale. On completion of clinical placement, educators assessed professional competencies considered integral to the appropriate practice of physiotherapy at entry level and rated the items according to behaviour descriptors. Following the pilot trial, refinements to the instrument (Table 5.6) were made based on qualitative and quantitative data collected during the trial. Phase Three (Chapters 6, 7 8, and 9) of this thesis is the field testing of the refined assessment instrument across a large and representative sample of students and their clinical educators in Australia and New Zealand. The first field test was designed to provide both qualitative and quantitative data for evaluation of the instrument.

This chapter describes the three stages of Field Test One. Stage 1; Preparation for Field Test One, Stage 2; during the field test, and Stage 3; on completion of testing. In addition this chapter presents the results of analyses of qualitative data collected during these stages. Chapter Seven focuses on the results of Rasch analysis of Field Test One quantitative data and summarises the modifications made to the APP instrument prior to the second and final field test.

6.2 Methods : preparation for Field Test One

6.2.1 Development of training resources

As described in Chapter Five, all participants agreed on the importance of training clinical educators and students in use of the assessment instrument. Themes relating to preparation of training material included accessibility, standardisation, time efficiency, variety in modes of training delivery, clarity and succinctness of information, definitions of

performance standards, role of formative and summative assessment, involvement of students in self assessment, exemplars of performance standards and guidelines for practise in use of the instrument. These issues were addressed when training resources were developed prior to field testing. To enable all participants involved in Field Test One to be adequately prepared and to maximise attendance, the researchers provided standardised training to educators and students using a number of approaches.

6.2.1.1 APP Resource Manual

A training package to support clinical educators in the application of the instrument was developed prior to Field Test One. The development of the draft training manual was guided by the principles of androgogy (adult learning) including the following:

- Adult learners are pressed for time
- Adult learners are goal oriented
- Adult learners bring previous knowledge and experience
- Adult learners are autonomous and self directed
- Adult learners are relevancy oriented
- Adults are practical, focusing on the aspects of a lesson most useful to them in their work (Knowles, Holton, & Swanson, 2005).

Additionally, development of the manual incorporated instructional design principles tailored to distance education (Carliner, 2003). These design principles targeted the following features:

- Flexibility and accessibility : to accommodate a broad range of individual abilities, preferences, educational needs, diversity of location and access to on-line technology;
- User friendly technology : the design allows for a variety of individuals of differing skill levels to use the information with minimal physical and cognitive effort
- Simplicity: unnecessary complexity is avoided

The design and contents of the Field Test One resource manual were developed with consideration of the stated needs of clinicians that were expressed during focus group discussions held prior to and following the pilot trial. The educators requested that the

resource manual provide easy-to-find answers and guidance, address frequently asked questions, be brief but comprehensive, and provide definitions of factors that affected the use of the instrument. The support material needed to be freely available in both hard and electronic copy to users of the APP.

6.2.1.2 Training workshops

In addition to the development of the training manual, two workshops were developed that combined an update on instrument development and pilot trial results with discussion of the instrument (APP version 3) and its application in the proposed field test. One presentation was developed for use with students and the other for clinical educators. All information presented in the workshops was contained in the training manual.

The workshops were designed for experiential learning and were guided by the work of Bourner et al (1994) and Race (2002). The basis of experiential learning is that the learner is directly involved in an event and then draws conclusions from it; hence the learner is an active participant rather than passive recipient (Kolb, 1984).

6.2.2 Development of demographic data forms and feedback questionnaires

A clinical educator and student demographic data form and feedback questionnaire were developed prior to Field Test One. The demographic form collected data required for investigation of factors that might bias measurements (differential item function) and included the following factors:

- Field of practice (neurosciences, musculoskeletal sciences, cardiorespiratory sciences)
- Student age and gender
- Clinical educator age and gender
- Clinical educator self-rated level of experience as an educator
- Clinical educator completion of clinical educator workshop
- Facility type, e.g. public hospital, private hospital, community health centre
- University attended by students being assessed

The questionnaire requested feedback on the instrument and training methods, as well as time to complete assessment and the acceptability of the APP in the clinical context. Questions dealt with users' experience of the assessment instrument, its perceived face validity, and factors that might affect its reliable and valid use (MacLellan, 2001). The questionnaire was designed according to principles that have been argued to maximise effectiveness of consumer evaluation surveys (Tourangeau, Rips, & Rasinski, 2000). Respondents were asked to rate their agreement to ten statements using a 5 point scale (1 = Strongly Disagree and 5= Strongly Agree) so the strength of the agreement could be indicated. Response options allowed a neutral opinion of '3'. Statements were worded so that the meaning was unambiguous and one issue only was addressed by each statement. On the educator questionnaire, two open ended questions were included to solicit additional feedback that was not targeted with the ten statements: "Were there any additional performance indicators that you consider could be added to the APP?" and "Do you have any additional comments on the APP and performance indicators". On the student questionnaire four open ended questions were included: "Overall, I consider the scores I received for each of the 20 items were a fair indication of my performance, if not please comment", "What needs to be done prior to each clinical unit to ensure students fully understand the role of the APP in assessment?", "What needs to be done prior to each clinical unit to ensure the clinical educators fully understand the role of the APP in assessment?" and "Do you have any additional comments on the APP and Performance Indicators?". Demographic forms and feedback questionnaires (Appendix 6.4) were completed by students and educators on completion of each clinical unit during Field Test One, returned by mail to the project manager (MDal). Data from demographic forms were matched with educator and student identification numbers, entered into a spreadsheet and deidentified.

6.2.3 Recruitment of participants

A broad approach to recruiting participants was employed to facilitate comprehensive and representative data collection and involvement of all stakeholders. All Australian and New Zealand physiotherapy programs were approached and provided with information briefly describing the APP research and requesting their support and interest in involvement in Field Test One. All 16 Australian and both New Zealand programs provided in-principal

support for the project. Logistically it was neither possible, nor considered imperative, to obtain ethical clearance and manage data collection from 16 programs. Nine were included, based on interest, timely commitment to participation and logistics. Five of the nine participating universities were involved in the original ALTC grant application (Table 6.1). Ethics clearance was obtained from the human ethics committee of each participating university (refer to Appendix 3.4). At each of the nine participating universities, the clinical education manager assumed the role of liaising with the research group for the field test. The education managers accepted responsibility for collection and return of educator and student consent forms to the research group.

Information on the project (provided in face to face meetings and in writing) was provided to students undertaking major clinical placements from each participating university. Following presentation of information on Field Test One, interested students were invited to sign forms consenting to the analysis of their deidentified assessment and questionnaire data. Students returned consent forms to a discrete mail box within each university department. This process was designed to limit the potential for student coercion into research participation (Appendix 6.1).

In addition to placement coordinators and students, the voluntary and informed participation of clinical educators was also targeted. Clinical educators who were assigned to the education and assessment of consenting students were sent an information sheet and consent form and invited to participate (Appendix 6.2). Assessment data were excluded from analysis if either the student or their clinical educator did not consent to participation in the research. Participants were advised that all data would be de-identified prior to data analysis.

Of the nine programs participating in Field Test One, six programs agreed to use the APP in parallel with their current university specific clinical assessment form. Three programs elected to use the APP as the sole assessment instrument (Table 6.1). In the six programs where the APP was used in addition to usual grading procedures, students were advised that the scores provided by the clinical educators on the APP instrument would not be used in establishing their grade for the clinical unit.

Table 6.1: Field Test One: Participating universities

Participating University	APP only	APP & university specific assessment instrument	Collaborator in the ALTC grant
Griffith University, Qld		X	X
University of Sydney, NSW		X	X
University of Otago, NZ		X	
Auckland University of Technology, NZ		X	
Curtin University, WA		X	X
Charles Sturt University, NSW		X	
Monash University, Vic	X		X
La Trobe University, Vic	X		X
James Cook University, Qld	X		

6.2.3.1 Training of participants

Visits to universities and clinical agencies across Australia were undertaken by the project team before and during Field Test One to disseminate information about the research results to date and for training clinical educators and students in the use of the instrument. Where possible visits were timed to coincide with regular facility-based clinical education workshop activities to limit the organisational burden and maximise attendance. All clinical educators received training through workshop attendance and/or access to the APP resource manual. Compulsory workshop attendance for all clinical educators participating in Field Test One was not feasible in the authentic clinical education environment where uniform face to face training opportunities are constrained by geographical, workload and financial considerations. While standardisation of delivery methods for training was not feasible, detailed mapping of workshop and resource manual content was designed to standardise information provided to all participants.

Students were educated in the assessment process and use of the APP instrument by a member of the research group or the clinical education manager at each university using a standardised presentation developed by the research group and information about the APP was included in the student clinical education manual for each university.

6.3 Methods Stage Two: during Field Test One

6.3.1 Field Test One procedure – during the clinical education unit

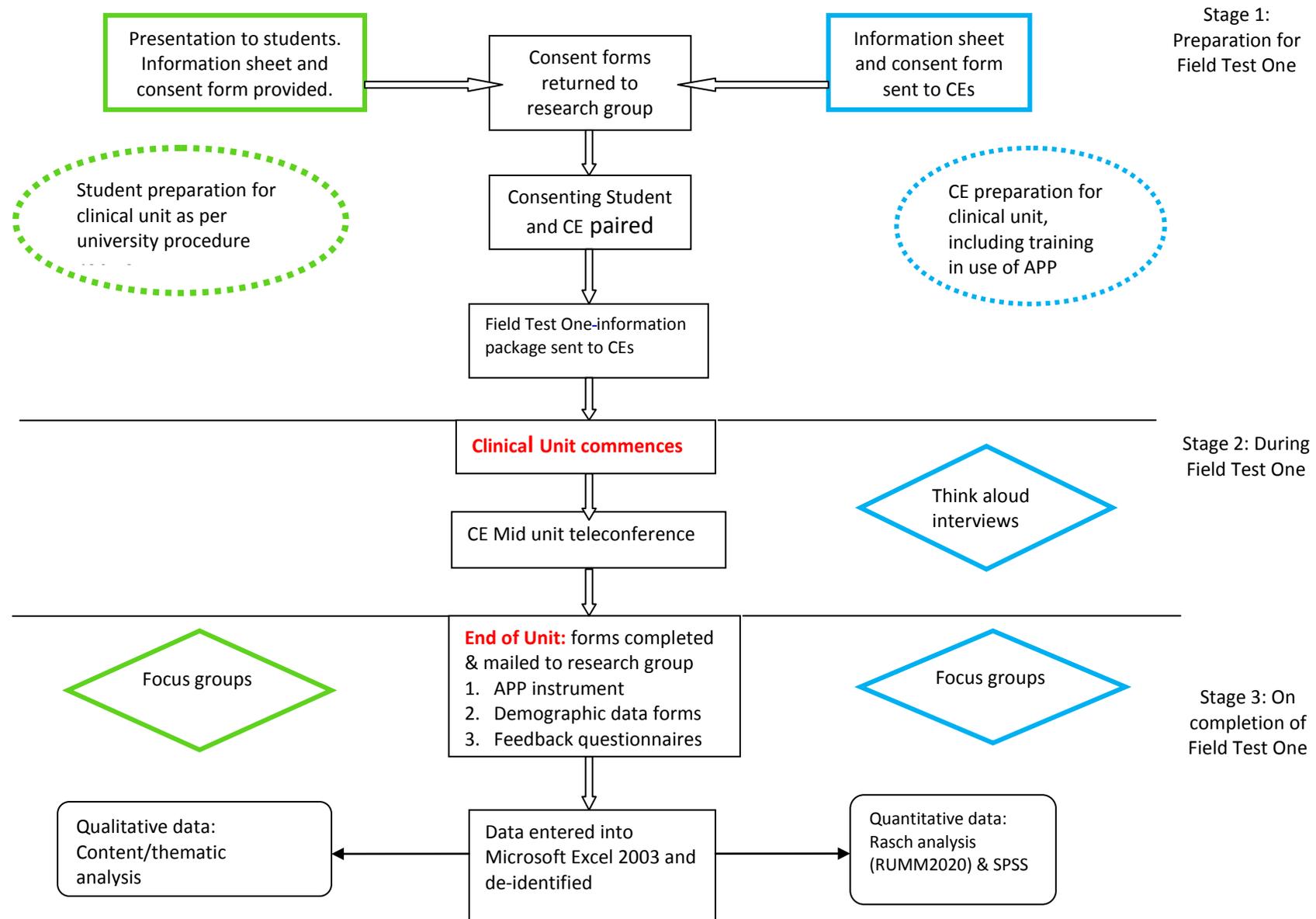
Each participating clinical educator received a field test package just prior to commencement of a clinical placement. The package contained the resource manual, a copy of the APP (version 3) instrument for each student, a clinical educator and student demographic data form and feedback questionnaire. A reply paid envelope was provided to facilitate return of completed forms.

Students from the nine participating universities completed clinical units ranging in length from four to six weeks with all students engaged full time in clinical education. The clinical units represented the major areas of physiotherapy practice and included musculoskeletal, cardiorespiratory, neurological, paediatric and gerontological physiotherapy. For the six programs using the APP in parallel with a current university-specific form, the educators were instructed to assess the student's performance using the APP at the end of the clinical unit prior to completing the required university assessment documents. In these six programs, students did not view the completed APP instrument. Mid unit formative feedback was provided using the current university-specific form. In the three university programs where the APP was the sole assessment instrument, the educators completed the APP at mid and end of the unit. During end of unit summative assessment, students viewed the completed APP instrument.

6.3.2 Teleconferences with clinical educators

On commencement of each clinical unit, educators were emailed an invitation to link into a 30 minute teleconference that was scheduled to take place mid way during the unit. The email outlined the purpose of the teleconference, time, date and conference call access phone numbers. All teleconferences were conducted by one member of the research group (MDal). Attendance at a teleconference was not compulsory but recommended. The teleconferences provided support for the clinical educators and assisted in standardising information about the use of the APP. Any issues raised during teleconferences were recorded in an APP issues register and collated at the end of Field Test One using the coding guide presented in Table 6.5. This enabled the research group to determine the issues that

concerned the educators and the frequency and occurrence of consensus of participants regarding each issue. A summary of Field Test One procedures is presented in Figure 6.1.



(Legend: CE= clinical educator; APP= Assessment of Physiotherapy Practice Instrument)

Figure 6.1: Flow chart of Field Test One procedure

6.3.3 Think aloud interviews

An important step in compiling validity evidence during instrument development was to engage in a detailed analysis of individual responses while the educators were completing the instrument (think aloud interviews) or just after completing an assessment in an exit interview (American Educational Research Association, 1999; Wilson, 2005).

Think aloud interviews have been recommended as a method for exploring what educators are thinking when administering the assessment instrument (section 1.4.4). Data from these interviews enable a view of aspects of an instrument that are ambiguous or inconsistently interpreted and are valuable in refining and validating assessments during development. Interviews enable in depth exploration of responses, facilitating high quality feedback that may not be possible with written survey responses completed at varying time points following assessment instrument use. Drennan (2003), Willis (2005) and Lam and Kolic (2008) recommended that qualitative data is collected through interviews to understand how and why people respond to items and scales as they do.

6.3.3.1 Participants

A subgroup of clinical educators engaged in data collection were invited to participate in think aloud interviews during which they were asked by the researcher to 'think aloud' as they scored each item and completed the global rating scale. Purposive sampling was used to achieve the input of broadly representative and qualified participants. Recruitment was designed to optimise representation of all stakeholders by location (metropolitan, regional/rural and remote), clinical area of practice, years of experience as a clinical educator/supervisor or manager, organization (private, public, hospital based, community based and non-government) and whether the APP was being used in parallel or as the sole assessment instrument.

6.3.3.2 Duration and site selection

Each interview was scheduled for one hour and arranged at a time and location to suit the participating educator.

6.3.3.3 Moderator

The principal moderator for each interview was a research assistant with expertise in interview methods. The research assistant was not a physiotherapist and had no vested

interest in the outcomes of the research, countering the potential for moderator bias. A second moderator handled the logistics, such as refreshments, set up of the interview venue, completion of consent forms, recording interviews using digital recording devices and communication with participants following interviews.

6.3.3.4 Recruitment procedure

A potential participant pool was collated. This pool was based on knowledge by the research team of individuals and organizations who take physiotherapy students for clinical education placements or who managed staff responsible for clinical education. For logistical reasons, the pool was limited to Australian clinical educators.

Initial contact with potential participants was made by email and/or phone and discussion included background information about the purpose of the interviews and an invitation to participate. Every invited clinical educator agreed to participate in the 'think aloud' interviews. Participants were sent an information sheet; a consent form and an outline of the procedure of the interview (see Appendix 6.2). Follow-up phone calls were made two to three days after initial contact to confirm that those who had agreed to participate had received the material that described the proposed content of the think aloud interviews and to provide additional information requested by the participants. Interviews were scheduled for the last two days of the final week of the clinical placement when the clinical educators would normally be completing the assessment instrument prior to the summative feedback session with the student.

6.3.3.5 Think aloud interview protocol

If any participant had not signed a consent form, this was finalised prior to the interview commencing. The proposed structure of the interview was outlined to participants. They were advised that the interview would be recorded, transcribed and de-identified, that the data were confidential and that they were free to cease the interview at any time. Clinical educators were instructed to 'think aloud' as they completed the APP instrument. The interviewer used questions (Table 6.2) to prompt disclosure of cognitive processes that educators used in arriving at a rating for each item and for grading using the global rating scale.

6.3.3.6 Questions

A list of possible questions for use in each interview was compiled based on feedback from the focus groups held following completion of the pilot trial. The questions were designed to investigate the level of understanding and interpretation of item content, performance indicators, item scoring scale, passing standard, and the global rating scale. An interview recording form was developed specifically for these interviews, a sample of which is presented in Table 6.2.

Table 6.2: Sample think aloud interview data collection form

Interviewer:	Date:
Interviewer ID:	Time:
Demographic form completed: Yes No	Location:
Clinical area:	Consent signed: Yes No
Q1: Are there any items you do not understand? Possible additional questions: Which one? Why?	Notes
Q2: Did the information in the resource manual assist you when completing the APP?	Notes
Q3: Can you explain your understanding of the rating scale and how you use it to score each item? Possible additional questions: Do you use a rating of 0? Do you use a rating of 4?	Notes
Q4. Can you please outline what you consider is a passing performance for a student in this unit?	Notes
Q5. (If the educator circles not assessed for an item) Why have you chosen n/a for that item?	Notes
Summary of issues	

6.3.3.7 Data management

Two audio recorders were used to capture interview content in case one failed. Recorded interviews were transcribed by a second moderator and compared to the notes on the interview data collection form. This cross check of data was performed to ensure all issues raised during the interviews were comprehensively recorded on the interview forms. Within seven days of each interview, participants were sent a de-identified copy of the interview transcript and asked to confirm whether this accurately reflected the interview. If not, they were invited to request amendments. Once the interview recordings had been transcribed, confirmed and de-identified, audio tapes were erased.

6.3.3.8 Data analysis of think aloud interviews

As described in Chapter Three (section 3.3.1.7), the method of analysis chosen for this study was a synthesis of qualitative methods of thematic analysis. An a priori template of codes was developed based on theory and prior knowledge of issues relating to assessment of clinical performance. This was integrated with data driven codes (Boyatzis, 1998; Crabtree & Miller, 1999; Fereday & Muir-Cochrane, 2006). The content analysis of all transcripts and field notes was conducted independently by two reviewers and followed five steps (Barbour, 2005; Bogdan & Biklin, 1998; Hsieh & Shannon, 2005; Kitzinger, 1995; Melia, 1997; D. L. Morgan, 1988; Sandelowski, 2009):

1. Review of the transcript and observer notes
2. Creation of a coding guide using the interview questions and theory underpinning clinical education assessment practices
3. Application of the coding guide to the interview data
4. Identifying themes in the data, including those not covered by the questions
5. Interpreting the data

6.4 Methods Stage 3: on completion of Field Test One

6.4.1 Data management and analysis

On completion of each placement completed forms were returned by mail to the project manager (MDal) and entered into Microsoft Excel 2003. All data were de-identified once entered into spread sheets for statistical analysis and names of educators, students and physiotherapy programs were replaced by codes. Data were checked for accuracy and the links between names and codes were then permanently destroyed. Data analyses for Field Test One were performed using SPSS 14.0 (SPSS Inc.) and RUMM2020 software (Andrich, et al., 2003) for Rasch analysis (refer to Chapter Seven).

6.4.2 Focus groups conducted following Field Test One

As discussed in Chapter Three (section 3.3) and Chapter Five (section 5.2), the aims of the focus groups, presentations and discussions conducted during and after Field Test One were to:

- disseminate information about the project and inform stakeholders;

- gather comprehensive feedback from stakeholders
- engage the physiotherapy profession in the participatory nature of the research.

This approach also enabled triangulation and reinforcement of decisions based on qualitative data obtained from multiple sources.

Four focus groups (two for clinical educators, two for students) were conducted to gather feedback on the draft items, the proposed rating scale and the performance indicators using methods previously described (Chapter Three, section 3.3.1 and Chapter Five, section 5.2). Questions introduced in the educator focus groups covered the items and performance indicators (content, wording and clarity), scale (size, format, pass level, understanding of levels of performance), layout of the instrument, and training in the use of the APP. The student group questions focussed on their experience of the instrument and included the following:

- What do you think are the strengths and weaknesses of the APP instrument?
- What aspects of current assessments would you keep?
- What aspects of the current assessments would you change?
- What is missing from the current assessments?
- Overall, do you consider the scores you received for each of the 20 items were a fair indication of your performance
- What needs to be done prior to each clinical unit to ensure students fully understand the role of the APP in assessment?

Qualitative data from all sources (training workshops, teleconferences, think aloud interviews, focus groups, and questionnaires) were analysed individually and then collated and sorted by topic and content. Issues that arose consistently and where there was clear consensus that change was required, were dealt with immediately; issues where divergent views were held by stakeholders e.g., number of categories on the rating scale, were reviewed with consideration of both quantitative and qualitative data. Issues that arose infrequently were dealt with on a case by case basis. The research group determined that requests for changes to the APP instrument that contrasted with recommendations for the development of high quality instruments would not be agreed to unless a sufficiently robust

counter argument was made. Analysis of both the qualitative (Chapter Six) and quantitative (Chapter Seven) data provided the basis for refinement of the APP after Field Test One and before commencement of Field Test Two.

6.5 Results: Qualitative evaluation Field Test One.

Ethics approval for all stages of this section of the research was obtained from the Human Ethics Committees of Griffith and Monash Universities and from the Human Ethics Committees of each university where a physiotherapy program leader had agreed to participate in data collection in either the pilot trial or any of the subsequent field tests (Appendix 3.4).

6.5.1 Development of training resources

The training package consisted of a brief, easy to read manual providing a broad overview of relevant information on issues relevant to assessment of physiotherapy students using the APP. Questions and answers regarding appropriate use of the APP raised by educators or students during focus groups and training activities were included in the manual as a 'frequently asked questions' section. The complete resource manual is provided for review in Appendix 6.3. All clinical educators involved in Field Test One data collection received a copy of the training manual. The manual assisted in standardising information about the use of the APP.

6.5.2 Demographic data form and feedback questionnaire

A student and clinical educator demographic data form and feedback questionnaire were developed and included in the Field Test One resource package sent to all clinical educators (Appendix 6.4)

6.5.3 Participant characteristics

In Field Test One a total of 747 completed APP assessments from 529 students were returned by 355 clinical educators. Physiotherapy programs delivered by nine universities participated in the field test. The 529 students were completing clinical placements of four to six weeks duration during the last 18 months of an entry level physiotherapy degree.

Table 6.3 summarises participant characteristics.

Table 6.3: Field Test One participant and placement characteristics

9 Universities: Australia=7, NZ=2	Characteristics
Student (n=529)	
Age (years) <i>Mean ± SD</i>	22.6 ± 3.4
Age range (years)	19 - 50
Gender %	68% F
Missing data	8%
CE (n=355)	
Age (years) <i>Mean ± SD</i>	34.4 ± 8.9
Age range (years)	22 - 67
Gender	78%F
Years of experience as CE <i>Mean ± SD</i>	7.1 ± 6.18
Range (years of experience)	0 – 36
% CEs attending training as an educator (prior to field test training)	60% yes, 20% no, (20% missing)
Self rating of experience as a clinical educator <i>Mean ± SD</i> (1= no experience – 5= very experienced)	3.43 ± 1.07
Missing data	11%
Time taken to complete APP (mins) <i>Mean ± SD</i>,	21.65 ± 13.3
Range (mins)	5 - 85
Clinical area (% of time spent in clinical area during unit)	
Cardiorespiratory physiotherapy	21.0
Neurological physiotherapy	22.0
Musculoskeletal physiotherapy	46.0
Paediatric physiotherapy	5.0
Speciality units eg., spinal injuries, burns, women’s health, oncology, mental health, hand therapy, plastic surgery	4.5
Missing data	1.5
Patient/client age group (%)	
Children (0-12 years)	5.5
Adolescents(13-20 years)	5.0
Adults (21-65 years)	37.0
Older people (> 65 years)	49.0
Missing data	3.5
Type of facility (%) n=535	
Public hospital	59
Private hospital	7
Community based services	9
Private practice	5
Non-government organisation	5
Missing data	15
University Program (% of completed assessments n=747)	
La Trobe	29.9
Monash	22.8
Griffith	18.0
James Cook	8.3
University of Sydney	5.4
Curtin	9.6
Otago	1.3
Charles Sturt	3.0
Auckland University of Technology	1.0
Missing data	0.7

6.5.4 Participant training: Workshops, and teleconferences

A total of 296 clinical educators attended 27 workshops held across Australia and New Zealand as part of training for clinical educators involved in Field Test One (Table 6.4). All students received information regarding the research and were prepared for each clinical unit by a member of the research group or by a university clinical education manager. In addition, during the clinical units, 85 educators attended 10 teleconferences.

Table 6.4: Field Test One clinical educator training

Locations	No. of participants
Victoria	
Angliss Hospital, Box Hill Hospital, Maroondah Hospital, The Alfred Hospital, Caulfield General Medical Centre, MacKellar Centre, Northern Health Network, Monash University Peninsula Campus and Gippsland Campus	70
Tasmania	
Launceston Hospital	5
Western Australia	
Royal Perth Hospital Perth and Shenton Park Campuses, Charles Gairdner Hospital	31
New South Wales	
Westmead Hospital, Prince of Wales Hospital, University of Sydney clinical educators	55
Queensland	
<u>Brisbane</u> : Royal Brisbane Hospital, Royal Children's Hospital, Paediatric Statewide Rehabilitation Service, Princess Alexandra Hospital, The Prince Charles Hospital, QEII Hospital, Bayside health service district, Redlands Hospital, Royal Children's Hospital, Gait Laboratory (teleconference), Interdisciplinary community rehabilitation therapists meeting	88
<u>Far North Queensland</u> : Townsville Hospital including videoconference to surrounding districts (Cairns, Mackay, Proserpine, Mt Isa)	23
New Zealand	
Otago Clinical Educators, Auckland University of Technology	24
Australia and New Zealand	
Mid unit clinical educator Teleconferences (n=10)	85

6.5.5 Think aloud interviews

Nine clinical educators representing a broad cross section of clinical areas and health facilities were invited and agreed to participate in think aloud interviews. As described previously in Chapter Three, section 3.3.1.7, after reviewing the transcript and observer

notes, a coding guide was created using the focus group questions and a priori knowledge of clinical education and assessment practice issues (Table 6.5). Application of the coding guide to the interview data enabled identification of themes and assisted interpretation of the data.

Table 6.5: Coding guide for content analysis of think aloud interviews

Code	Content
It	Items: content, wording and clarity of intent
Sc	Scale: size, format, wording
Pass	Pass standard: passing performance,
PIs	Performance Indicators: perceived utility, number, content, clarity of intent, wording, suggestions for additional PIs
IFor	Instrument format: layout of instrument, perceived utility, suggestions for improvement
Tr	Training in the use of the APP: requirements of a training package,
Fback	Use of the instrument in providing feedback
Other	Other key words, ideas, themes

Themes arising from think aloud interviews are summarised below.

6.5.5.1 Items

Item 8 (selects appropriate methods for measurement of relevant health indicators) and item 13 (collaborates with patient/client to select appropriate intervention) were identified by the interviewees as the most difficult to assess. The wording of item 8 created confusion for educators. Some participants thought this item was meant to assess if the student used a relevant health outcome instrument, for example the Oswestry Low Back Pain questionnaire. Other educators thought this item was asking if the student had set meaningful goals to measure the patient's outcome following treatment, for example, the patient can now brush their hair. Item 13 appeared to cause confusion as some participants interpreted this item to mean that the patient had to nominate what intervention they wanted rather than this being a collaborative decision reached after discussion between therapist and patient. Item 20 (risk management) was understood, however a number of participants considered this item may be better assessed using a pass/fail scoring system or preferred that the term 'safety' rather than 'risk management' had been used. The easiest items to score were the first five items relating to professional behaviour and verbal

communication as participants had a clear concept of the expected behaviours for these items and how to rate them.

6.5.5.2 Scoring system

Some participants considered the five point rating scale may be improved if one or two additional rating categories were available as indicated in the following quote,

“I felt the 0-4 rating scales was little “narrow” for example I often wanted to put 3.5 as the student was achieving better than a 3 (good) so they got a 4 (excellent). An extra 1 option or a more graduated scale could be better.”

On the other hand, the majority of educators agreed that more rating categories had potential to confuse educators and create unnecessary anxiety when rating the items for summative assessment. There was no reported aversion to the use of 0 or 4 by the participants. Rating of an item as 0 mainly occurred at mid unit. Only students who subsequently failed the clinical unit received a zero rating for an item during end of unit summative assessment.

A number of participants completed the global rating scale first, and then rated each item while others sequentially rated items from 1 to 20, completing the global rating scale last. All participants were comfortable with using the global rating scale. Several participants reported that the GRS allowed them to rate a student’s performance overall as good or excellent while still be able to score individual items at a level below this, for example, at a passing level (rating 2). Two participants reported scoring items they considered “the most important” first. These were items 10 and 11 (interpreting assessment findings and prioritising patient problems).

6.5.5.3 Entry level passing standard

The use of entry level competence as the passing standard (rating 2) for each item was understood and used by all participants during the clinical unit in which the think aloud interviews were conducted. However, several participants reported that previously they had used an alternative model of grading students against ‘the expected competency during the

first practice block in third year’ or ‘the expected competency during the last practice block in fourth year’. They reported recognising that this created individually constructed and unregulated assessment targets, and limited the opportunity for discussion regarding what entry level standard should look like and how to best support all students to achieve desirable standards of performance. The shift to entry-level competencies as the benchmark against which to judge student performance required considerable cognitive effort for those participants accustomed to an ‘experience based’ standard.

6.5.5.4 Performance indicators

Educator use of the performance indicators during the think aloud interviews varied. Several participants relied heavily on the indicators to guide their rating of each item, whereas other interviewees only referred to the indicators for those items where they were unsure of the expected behaviours. The consensus of the participants was that the indicators assisted with the final rating of an item when they were deciding between scores. The indicators were considered invaluable when providing specific feedback to students on their performance, as they reduced the cognitive effort needed to *“find the right words to say to the student to describe what they were or were not doing correctly”*.

6.5.5.5 Use of not assessed (n/a)

Two participants had interpreted the n/a on the rating scale to mean ‘not applicable’ to this clinical unit. These two educators worked in the area of outpatient physiotherapy and had used n/a for item 18 (undertakes discharge planning). Their rationale was that the patients they were treating were not inpatients i.e., patients admitted to the hospital and hence could not be discharged.

6.5.5.6 APP clinical educator resource manual

All interviewees had reviewed the manual prior to the commencement of the clinical unit and there was consensus on the utility of the information provided. Participants requested more information in relation to biases in assessment and strategies to overcome these and for the frequently asked questions section to be expanded. Evidence of the influence of the

resource manual information on clinician thinking was reflected in the language used during the think aloud interviews. On numerous occasions participants used the wording provided in the manual when describing why they had rated an item at a particular level, and referred to the performance indicators frequently to explain their decisions on student performance.

6.5.6 Feedback questionnaires

Two hundred and forty-nine (70%) clinical educators and 243 (50%) students returned a feedback questionnaire. Questionnaires were returned by educators and students from all nine universities, by students across year levels, and from a representative spectrum of placement and facility types. The results of the questionnaires are presented in Tables 6.6 and 6.7

Table 6.6: Clinical educator feedback on APP

Question	Mean ± SD	Median
Confident using 0 – 4 rating scale	3.96 ±0.66	4.0
Confident using Global Rating Scale	4.0±0.76	4.0
APP practical in the clinical environment	4.1±0.6	4.0
Performance Indicators (PIs) useful	4.1±0.7	4.0
PIs easy to understand	4.1±0.6	4.0
Time taken to complete APP acceptable	4.2±0.7	4.0
Beginning practitioner definition helpful	4.0±0.8	4.0
Scoring rules helpful	4.1±0.7	4.0
Resource manual information on how to complete the APP was comprehensive	4.2±0.7	4.0
Preference for on-line version	3.3± 1.0	3.0

Note: each item scored on agreement scale 1=Strongly Disagree, 2=Disagree, 3=Undecided, 4=Agree, 5=Strongly Agree

In relation to the final item on the questionnaire, ‘In the future, I would prefer to complete the APP on-line rather than posting/faxing hard copies’, 31% of clinical educators disagreed with this statement, 38% were neutral and 24% agreed with this statement.

Table 6.7: Student feedback on APP

Question	Mean \pm SD	Median
Confident CE used 0-4 scale correctly	4.1 \pm 1.0	4.0
PIs useful to assess own performance	3.9 \pm 0.7	4.0
Scoring rules appropriate	4.0 \pm 0.9	4.0
Entry level performance (pass) was clear to me	4.0 \pm 0.9	4.0
Items easy to understand	3.9 \pm 0.7	4.0
APP practical for use in clinical environment	4.1 \pm 0.7	4.0
Performance required to score 4 was clear to me	3.9 \pm 1.0	4.0
Information about APP prior to unit was adequate	4.1 \pm 0.7	4.0
Rating on GRS was a fair indication of my performance	4.0 \pm 1.3	3.0
Rating on 20 items were a fair indication of my performance	3.9 \pm 1.2	3.0

Note: each item scored on agreement scale 1=Strongly Disagree, 2=Disagree, 3=Undecided, 4=Agree, 5=Strongly Agree

The four open ended questions on the student questionnaire provided additional feedback with examples provided below:

Overall, I consider the scores I received for each of the 20 items were a fair indication of my performance. If not please comment

- *“I had an incident where one patient in the last week was standing and trying to adjust his crutches while standing with me – this was unsafe and I think she marked me low at the end of the unit because of this”*
- *“I felt I did much better than my results reflected, my educator missed/didn’t see a lot of the work I was more proud of”*
- *“I feel that the educator had an aversion to giving 4’s. He told me he doesn’t really give 4’s. But I feel if I deserve one I should get it”*
- *“Overall, yes I think the marks I received were very fair”*

What needs to be done prior to each clinical unit to ensure students fully understand the role of the APP in assessment?

- *“Leave as is”*
- *“Explain to students the effectiveness and importance of going through the detailed APP in the booklet while they complete their self reflection sheet. It was very useful for my own expectations of myself”.*

- *"I don't think anything needs to change. I understood it"*
- *"Nothing, it's good"*
- *"Tell them (students) to read the clinical placement manual!"*
- *"There is plenty of information if we need to find out more"*

What needs to be done prior to each clinical unit to ensure the clinical educators fully understand the role of the APP in assessment?

- *"Make sure they attend workshops and learn how to complete it properly"*
- *"Not sure, but they do need training so they don't ask us (the students) questions on what to do with the form"*

Do you have any additional comments on the APP and Performance Indicators?

- *"Adequate – is that satisfactory / pass? Better than pass?
Good – what is good, good is a poor description of performance
Excellent – what is excellent? Perfect or excellent for a student or in professional in the field?"*
- *"No"*
- *"Students need to read it!"*
- *"Should be more than 4 rating options"*
- *"I appreciated the prac and the level of feedback received was helped by the performance indicators"*
- *"Make scale 0-5 with 2 a pass mark"*
- *"Evidence based practice – is sometimes difficult to accomplish on placement, as it expected that you have the most up to date information on all aspects of physiotherapy"*
- *"Good source of feedback – very useful"*
- *"My educator seemed to understand the APP indicators well, hopefully others did too"*

6.5.7 Demographic form: Clinical educator experience

On the clinical educator demographic form the educators were requested to record how many years they had worked as a clinical educator and also to rate their perceived level of experience as an educator on a 5 point scale ranging from no previous experience to very experienced. The relationship between the years of experience as a clinical educator and the self rated level of experience as an educator was investigated using Spearman's rank order correlation coefficient. There was a strong, positive correlation between the two variables [$r=.72$, $n=243$, $p<.0005$], with high levels of self rated experience as an educator associated with greater number of years working as a clinical educator.

6.5.8 Focus groups conducted following Field Test One

Two focus groups for clinical educators ($n=20$), and two groups for students ($n=16$) were conducted on completion of Field Test One. Each focus group included a broad cross section of participants. Participants confirmed that the transcripts accurately reflected the focus group discussion and no requests for amendments to the transcripts were made.

6.5.8.1 Clinical educators

Themes arising from the clinical educator focus groups are summarised below.

- 1) Clinical educators provided positive feedback on the APP layout (practicality and comprehensiveness).
- 2) The one page format was viewed very positively.
- 3) Clinical educators agreed the items were an adequate representation of the competencies required of new graduate physiotherapists and were transparent for both educator and student.
- 4) Weighting of areas of practice by number of items was of some concern, for example, "communication and professional behaviours" have 6 items, while "intervention" has 5. Some clinicians felt that this over-weighted professional behaviours relative to practical skills.
- 5) In this field test educators agreed that the concept of risk management rather than 'safety' (a term used previously in many assessment instruments) was more appropriate terminology.

6) Educators felt that the performance indicators were very useful especially as they were written as observable behaviours, assisting them in giving specific feedback to students on the areas of their performance that were adequate and those requiring improvement.

7) Similar to the pilot trial, participants continued to report some confusion over how to score item 19 (applies evidence based practice in patient/client care) in work-based practice.

8) Educators were of the opinion that the rating scale and scoring categories were reasonable, but there was some concern about students in their first clinical units being able to achieve a pass standard (score of 2) on items if the pass level is set at entry level/beginning physiotherapist. This was not a view shared by all educators.

9) Some educators felt that 3 passing categories (scores of 2, 3 and 4) are sufficient to be able to adequately assess the performance of students while others expressed a preference for additional rating categories. A comment by an educator provides the a summary of the focus groups' attitude toward the scoring system.

"I think standardising the marking process is an excellent initiative but I personally like a little more scope for marking and was not used to putting things in a more defined box, but I guess this is the whole point and perhaps it will make educators read the competency information and make a standardised decision."

11) Some educators considered students should have to obtain a minimum score of 2 on each of the 20 items to pass overall but this was not a view shared by all.

12) Educators considered the Global Rating Scale (GRS) to be a useful internal check on item scoring and valuable as an overall impression of student performance. However, during summative assessment educators appeared to be using two different benchmarks when scoring the GRS, with some using entry-level, and others rating the student globally on their performance relevant to their progress through the clinical program or against other students in the cohort. Participants considered the use of different scoring benchmarks may be due to lack of alignment between wording on the item rating scale and GRS wording implying that the GRS was to be scored differently to the items.

13) Educators thought the training manual was comprehensive, and the frequently asked questions section and information on avoiding rater bias were particularly helpful. Several

participants requested that more information on how to assess item 19, application of evidence based practice, and time management be included in the frequently asked questions section of the manual.

One educator's comment reflected the overall mood of both educator focus groups;

"Thank you for providing a fair, comprehensive, equitable tool. I feel one of its most valuable aspects is it enables a very clear method to communicate back to students their performance level. There is less opportunity for misunderstanding and students seem clear on what the educator's expectations are."

6.5.8.2 Students

The student focus groups provided different data to the clinical educator groups. The educator group discussion was detailed, focused and collegial, providing greater consensus on most issues. In contrast, as might be expected, the student responses were more personally oriented and often varied widely between students depending on personal experiences during a clinical unit.

As an example, when the question, "Overall, do you consider the scores you received for each of the 20 items were a fair indication of your performance?" was asked, one student's response was;

"I felt competent and the verbal feedback I received seemed to reinforce that and my marks were indicative of this",

while another student with the same clinical educator reported;

"I believe that my score and the other students' scores were well below what was deserved".

Similarly, when asked about the feedback they received, the responses varied from;

"the feedback I received was always spot on, and the way the educator gave the feedback was very helpful, made you want to try harder" to;

"I feel as if my educator based my end of unit marks and feedback off my first week's performance and that he didn't change his mind from then".

Overall consensus was reached by the student groups on the adequacy of the training they had received prior to commencement of the clinical unit and that the items and performance indicators were easy to understand, comprehensive and helpful. Student comments on the question, “Do you consider your educator understood the role of the APP in assessment?” was variable. Overall however, the majority of students considered their educator was well prepared for their role as an educator with an adequate understanding of assessment practices. Table 6.8 presents a summary of the student focus group findings.

Table 6.8: Summary student focus group results Field Test One

Target issue	Student focus group results
Items and domains of practice	<ul style="list-style-type: none"> • All students agreed item content was comprehensive • Item 19, evidence based practice, was not well understood by clinical educators
Rating scale	<ul style="list-style-type: none"> • Half of the student participants considered need for additional scoring categories ie., 6 or 7 categories rather than 5 • Often verbal feedback and scoring on items did not correlate. Verbal feedback often very positive but scoring low. • Occasional comments that different educators (in the one unit) had different ideas on performance required to score a 2, 3 or 4.
Feedback /reflection	<ul style="list-style-type: none"> • Often verbal feedback and scoring on items did not correlate. Verbal feedback positive but scoring low. • Some educators scored items as '1' at mid unit to make student work harder in the second half of the clinical unit • If more than one educator, feedback could vary between educators • Performance in the first week of the unit often influenced final unit score. • Receipt of a copy of mid unit feedback was variable. Overall consensus that educators could provide more regular feedback • Engagement of students in mid unit self reflection on performance was variable
Performance Indicators (PIs)	<ul style="list-style-type: none"> • All students considered PIs very useful to guide assessment particularly during mid unit formative feedback • All students considered PIs were comprehensive and easy to understand
Overall, I consider the scores I received for each of the 20 items were a fair indication of my performance	<ul style="list-style-type: none"> • Wide variety in responses from students, with some agreeing and others considering their marks were too low. No student considered their scores were too high.
Training for students	<ul style="list-style-type: none"> • Majority of students were satisfied with training provided on APP and assessment during clinical unit. • Presentation and information provided in clinical manuals was considered comprehensive, but students also reported variability in reading of information
Training of educators	<p>Wide variety of comments:</p> <ul style="list-style-type: none"> • Several participants felt educators needed to spend time outlining their expectations for the unit and giving specific examples of different levels of performance • Other students considered educators were well prepared • Training for educators should be compulsory • Sign-off on final APP scores by student and educator was variable

6.6 Discussion

During the first field test of the APP instrument, the opinions of a broad cross section of the physiotherapy profession were collected through a variety of approaches. This procedure allowed triangulation of data from multiple sources and enabled comprehensive representation of the profession to input during the process of instrument refinement. Overall, qualitative data demonstrated strong convergence in the opinions of educators and students regarding several aspects including comprehensive content coverage of items and performance indicators, ease of use of the instrument within the clinical context, importance of performance indicators in providing clear, well-targeted feedback on performance, role of training of all stakeholders in assessment processes, and the effectiveness of the resource manual in providing accessible information on assessment practices.

The qualitative data also highlighted several unresolved issues requiring further investigation. Educators and students considered additional scoring categories may assist the rating scale to be more effective in identifying different levels of performance. However, the educators were also cognisant that the results of the pilot trial (Chapter Four section 4.3) demonstrated that the current five point scale was adequately differentiating student performance and that additional categories could create indecision on appropriate item ratings for less experienced educators. Students considered a wider scale would allow an improvement in performance to be more obvious, but were also in agreement that scale use had to be standardised between educators.

The global rating scale was viewed positively by the educators providing them with an opportunity to provide their overall impressions of student performance in addition to individual item ratings. It became evident however that some educators were unsure of the standard against which the GRS was scored. In the think aloud interviews two educators scored the GRS in relation to the student's prior experience rather than against entry level performance as instructed. Focus group discussion revealed this behaviour did not appear to be a frequent occurrence but did warrant addressing in future training workshops and in

revision of the resource manuals. The desire to provide lower performing students with positive feedback appeared to be the driver for use of the GRS in this manner.

Students commented that they were unsure if their educator was following the instructions provided in the workshops and resource manual when completing their assessment.

Additionally, when a student had more than one educator, they reported considerable variability in feedback on performance from different educators. Research by Cross and Hicks (1997) suggested that failure to refer to guidelines in the manual on how to use and complete the instrument and/or application of implicit personal criteria by educators could contribute to student perceptions of assessment of work-based performance as inherently 'subjective' or 'unpredictable'. Ensuring constructive alignment between educator and student expectations of assessment of professional competence requires a multi-pronged approach including the following strategies: regular monitoring and evaluation of educator understanding and attitude towards assessment of professional competence, and comprehensive training programs to clarify expectations of student performance (Bursari, Scherpbier, van der Vleuten, & Essed, 2006; Notzer & Abramovitz, 2008).

Clinical educators all agreed that the time taken to complete the assessment instrument was acceptable with the mean completion time of 21 minutes. The longest reported time to complete was 85 minutes and focus group discussion indicated that this time included completing the University specific assessment instrument as well as the APP for the research group. Additionally educators reported that assessments of lower performing students took more time than those students easily achieving a pass. There was consensus that the instrument was easy to use and that the performance indicators provided assistance with item ratings.

Clear consensus was reached during educator focus groups on the importance of standardised training in assessment practices and educators were enthusiastic about the workshops they had attended as part of the research. These comments highlighted the importance educators placed on the role of training in providing effective work based learning for students. Educators also observed that using one assessment instrument for students from different universities promoted time efficiency in training and enabled problem solving of assessment issues common to all students. Use of a common instrument

allowed educators to standardise their expectations of performance, in particular in relation to the entry level passing standard. Both educators and students requested video exemplars of different levels of performance in future training resources.

The strong, positive correlation of high levels of self rated experience as an educator with greater number of years working as a clinical educator is of interest as it has been suggested that an educator's ability to make an accurate assessment of student performance is affected by the degree of previous experience. This experience provides the context within which comparative judgements are made (Alexander, 1996; Chapman, 1998). This aspect of assessment warrants further investigation in the quantitative data analysis of Field Test One results when differential item functioning for levels of educator clinical experience are examined.

Wass (2001b) recommended that the minimum appropriate standard be decided upon prior to use of an assessment instrument. The decision regarding how to set a pass standard for items, and for overall assessment, generated discussion at both clinical educator focus groups. In discussions about entry level/beginning physiotherapist standards (during 12 focus groups (ten in the pilot trial and two in Field Test One) there was clear consensus from participants that for consistent use of an instrument across programs, students should be judged on each item against the minimum performance targets expected of a novice (entry-level) practitioner. The focus group participants agreed that many students had only one clinical block within which to gain skills in core areas of practice (such as e.g. neurological rehabilitation). It was therefore essential that the pass standard at the end of that block was entry level practice. The target of clinical education was the acquisition of a minimum acceptable level of skill irrespective of when each clinical unit was completed. A target of entry level competence enabled ranking of students relative to a common standard.

During development of the instrument the final 20 items were agreed by all stakeholders to be of equal importance when assessing professional competence. Items were grouped under relevant domains of practice based on the APC professional standards. The listing of items under domain headings was to assist educators and students to map the assessment to the relevant standards and was not meant to imply differential importance of one item or

domain above another. Despite this explanation, some educators expressed some concern that the importance of the domains of practice were being differentially weighted due to the variable number of items in each domain, for example, professional behaviour had four items whereas treatment planning had two. Other educators commented that the attributes associated with professional behaviour increasingly may define quality practitioners and as such comprehensive assessment of this domain was essential. This difference in opinion between educators reflects the findings of Cross (2001) where clinicians demonstrated differing perceptions of the importance of some attributes of professional competence. The importance of professional behaviour in assessment of professional competence is also supported by the work of Papadakis et al.(2004) and Yates and James (2006) who showed that doctors with proved professional misconduct had poorer academic grades during their undergraduate degree and had more frequent reports of poor performance during clinical placements. They recommended that efforts to ensure a high level of professional behaviour begin during classes in medical school.

The comments from participants regarding the variable number of items in each domain were noted, however as all items were developed and agreed upon as essential by all stakeholders, it was not considered appropriate to alter the composition of the twenty items at this stage. Educators were advised this issue would be further investigated in quantitative analysis of data from Field Test One.

6.7 Chapter Summary

The overall aim of this research was to develop a valid assessment instrument to measure the workplace performance of student physiotherapists. An essential component of instrument development involves engagement with and input from all relevant stakeholders. During the first field test, focus groups, one to one interviews, surveys, workshops, email, and teleconferencing were employed to provide training, support assessors and gather input from as many stakeholders as possible. Chapter Seven presents the quantitative data analysis for Field Test One.

7. Chapter Seven: Field Test One - Quantitative evaluation

7.1 Introduction

The first field test was designed to provide both qualitative and quantitative data for evaluation of the newly developed APP instrument. Quantitative evaluation of the APP instrument was designed to investigate the nature of the scores when the instrument was used to assess undergraduate physiotherapy students in the authentic practice environment. It was of particular interest to assess the behaviour of scores for different items, to determine whether item scores provided evidence of measurements of a single underlying construct and to assess bias in scoring associated with test conditions. Collected data were analysed using the Rasch Measurement Model because it provides a sophisticated method for instrument development (section 2.2.4). Raw scores (0-4) gathered during Field Test One using the third version of the APP (Appendix 5.2), the sum of item scores (total scores) and the overall global rating for each student were examined. Analyses of qualitative data from Field Test One have been reported in Chapter Six, while this Chapter focuses on the results of Rasch analysis of Field Test One quantitative data.

7.2. Method

7.2.1 Participants – students and clinical educators

Students

Participants in the first field test were students enrolled in physiotherapy programs from universities in Australia and New Zealand. The third version of the APP (Appendix 5.2) was used to assess students during usual 4-6 week clinical placement blocks scheduled across one university semester in 2007. Students attended clinical placements on a full-time basis (32-40 hours/week).

Clinical educators

During clinical placements, students were supervised by clinical educators (graduate physiotherapists) in 1:1 to 1:4 educator:student ratios. Recruitment procedures (reported in full in Chapter Six sections 6.2.3) optimised representation of educators by location (metropolitan, regional/rural and remote), clinical area of practice, years of experience as a

clinical educator/supervisor or manager, organization (private, public, hospital based, community based and non-government).

7.2.2 Field testing procedure

The procedure for Field Test One has been described in detail in Chapter Six, sections 6.2 – 6.4.

7.2.3 Data management and analysis

Using methods developed and tested in the pilot trial (Chapter Four section 4.2), completed student assessment forms were returned to one of the researchers (MDal) by mail, entered into a spreadsheet and de-identified. The aim of data analysis was to investigate properties of the scores obtained using the instrument. Data analyses were performed using SPSS 14.0 (SPSS Inc.) and RUMM2020 software (Andrich, et al., 2003) for Rasch analysis. Data were coded as missing if an item was not scored on the 0 – 4 rating scale.

Tennant and Pallant (2006) p 1048 state that;

“when developing new polytomous scales, an exploratory factor analysis used a priori, with parallel analysis to indicate significant eigenvalues, should give early indications of any dimensionality issues prior to conducting Rasch analysis.”

Given this recommendation, Field Test One data were initially subjected to principal components analysis (PCA) using SPSS version 14 (SPSS Inc.) prior to conducting Rasch analysis.

7.2.3.1 Rasch analysis

Rasch models are developed on samples of data and sampling variation could, by chance alone, lead to the construction of a model that did not represent the typical behaviour of item responses. Validation of the model in an independent sample provides confidence in the model fit and the item hierarchy that emerges. In Field Test One, a large data set (n = 729) was available. Data were divided into two random samples, one (n=390) for model development and the other for model validation (n=340). The data were stratified and then randomised to optimise representation of completed APP instruments according to clinical area of the placement, level of student experience, facility type (hospital, non-government

agency, community health centre, private practice) and university program (undergraduate, graduate entry). Linacre (1994) argued that sample sizes greater than 243 enable adequate precision regardless of the targeting of the group or the distribution across the response options of each item.

Similar to the pilot trial (Chapter Four), Rasch analysis was conducted to investigate overall model fit, overall person fit and item fit, individual item fit, thresholds, targeting, person separation index (PSI) and local independence (dimensionality) (J. F. Pallant & Tennant, 2007). Explanations of these aspects of Rasch analysis have been presented previously in Chapter Four (section 4.2.4). Additionally in Field Test One, investigation of differential item functioning (DIF) was conducted. The Rasch measurement model tests for invariance of items across external group characteristics by differential item functioning (DIF) analysis, that is, a scale should function consistently irrespective of subgroups within the sample being assessed. For example male and female students with equal levels of the underlying construct being measured should not achieve systematically different scores. If this occurs across all class intervals (the group divided into sub-groups with low, medium and high levels of the trait) it represents the presence of uniform DIF. Non-uniform DIF occurs where there is inconsistency in the differences between the class intervals ie., differences vary across levels of the trait (Lai, et al., 2005; J. F. Pallant, 2007; Tennant & Conaghan, 2007; Teresi, 2001). An example of non-uniform DIF would be if among all students scoring lower on an item, females had higher scores, while in the group of students scoring higher overall on the same item, male students performed better. In this research the presence of DIF was investigated for the following variables;

- Clinical area (neurosciences, musculoskeletal sciences, cardiorespiratory sciences);
- Student age, gender and number of weeks of clinical experience;
- Clinical educator age, gender and experience;
- Facility type, e.g. public hospital, private hospital, community health centre; and
- University

7.3. Results

Ethics approval was obtained from the Human Ethics Committees of Griffith, La Trobe and Monash Universities and from the Human Ethics Committees of each university where a physiotherapy program leader had agreed to participate in data collection in either the pilot trial or any of the subsequent field tests (refer to Appendix 3.4).

7.3.1 Participant characteristics

In Field Test One a total of 742 APP completed assessments of 529 students were returned by 355 clinical educators. Thirteen of the 742 returned assessments were incomplete, leaving 729 (98%) assessments suitable for analysis. Nine university physiotherapy programs participated in the field test. The 529 students were completing clinical placements of 4 – 6 weeks duration during the last 18 months of their physiotherapy entry-level degree. Table 6.3 (Chapter Six) provides a summary of participant characteristics.

7.3.2 Characteristics of item and instrument scoring

Table 7.1 presents the descriptive statistics of the raw scores for each item, the raw total scores for the 20 summed item scores and the frequencies of use of each rating scale category for the 20 items on the 729 completed assessments.

The overall mean of the total APP scores for 729 completed assessments was 58.3/80 (72.8%) ($SD = 12.5$). The mean rating on the global rating scale (GRS) for 729 APP forms was 2.9 where 1= poor, 2= satisfactory, 3= good, 4= excellent. For students with 0-10 weeks of clinical experience prior to Field Test One, the mean APP score was 56.47/80 ($SD = 11.7$), for students with 10-20 weeks prior experience the APP mean score was 61.89/80 ($SD = 12.2$) and for 20-30 weeks, 67.21/80 ($SD = 10.5$). The GRS mean for 0-10 weeks was 2.8, 10-20 weeks, 3.0 and 20-30 weeks 3.3. The mean (SD) APP score for students receiving a GRS of poor was 29.89 (7.36) for satisfactory 47.33 (4.79), for good 57.76 (5.92) and excellent 72.17(5.05).

Table 7.1: Descriptive statistics Field Test One (n=729)

Item	N	Mean	Standard Error of Mean	SD	Rating 0		Rating 1		Rating 2		Rating 3		Rating 4		N/A	
					Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
1	729	3.32	.024	.69	0	0	5	0.7	82	11.2	317	43.3	325	44.6	1	0.13
2	728	3.23	.028	.81	1	0.1	1	0.1	112	15.1	287	38.3	327	43.6	0	0
3	729	3.38	.024	.71	1	0.1	1	0.1	62	8.5	295	39.9	369	49.9	0	0
4	729	3.09	.030	.86	1	0.1	32	4.4	139	19.0	284	39.0	272	37.4	5	0.68
5	728	3.09	.027	.78	0	0	24	3.3	120	16.5	350	48.0	234	32.1	0	0
6	729	3.12	.025	.72	0	0	10	1.3	123	16.7	370	50.8	225	30.9	2	0.27
7	729	2.98	.025	.72	2	0.2	19	2.6	133	18.3	417	57.3	158	21.7	0	0
8	727	2.79	.026	.75	3	0.4	30	4.1	193	26.5	393	53.9	109	15.0	5	0.68
9	728	2.82	.024	.70	2	0.2	20	2.7	185	25.4	420	57.6	102	13.9	2	0.27
10	727	2.68	.028	.82	3	0.5	54	7.4	220	30.2	344	47.2	106	14.5	0	0
11	729	2.74	.028	.80	5	0.6	42	5.7	202	27.5	369	50.6	111	15.1	0	0
12	729	2.67	.027	.79	3	0.5	41	5.6	238	32.6	355	48.6	92	12.4	7	0.96
13	727	2.79	.027	.79	6	0.8	29	4.0	199	27.2	369	50.6	125	17.1	6	0.82
14	729	2.93	.026	.76	2	0.2	28	3.7	150	20.5	393	53.9	157	21.3	1	0.13
15	729	2.86	.030	.86	6	0.8	40	5.5	175	24.0	334	45.8	174	23.9	3	0.41
16	728	2.83	.028	.82	4	0.6	38	5.3	182	24.9	359	49.2	145	19.9	0	0
17	729	2.72	.029	.83	6	0.8	44	6.0	219	30.1	339	46.5	121	16.6	1	0.13
18	717	2.74	.030	.85	5	0.6	50	6.7	200	27.3	339	46.0	123	16.6	22	3.01
19	720	2.73	.030	.87	13	1.8	42	5.6	196	26.7	348	47.7	121	16.5	32	4.38
20	728	3.00	.029	.83	4	0.5	30	4.0	143	19.4	340	46.6	211	28.9	8	1.09
GRS	728	2.90	.027	.75	Not applicable to GRS		Poor		Satisfactory		Good		Excellent		Not applicable to GRS	
							14	2.0	159	22.0	360	49.3	195	26.7		
Tot. score for 20 items		58.29 /80	0.46	12.54	Range of total raw scores for 20 items: minimum=11; maximum=80											

Legend: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies Evidence based practice 20 = assesses risk; N/A= not assessed; SD= standard deviation; N=number; GRS=global rating

The data presented in Table 7.1 show that the total score for the 20 items ranged from 11 – 80 highlighting a wide spread of scores. Item 18 (undertakes discharge planning) and item 19 (applies evidence based practice in patient care) were the items most frequently not scored. The missing data rate for item 18 was (1.6%), item 19 (1.2%) and the overall missing data rate was 0.20% (30 items not scored out of a possible 14,580 item scores). The frequency of use of not assessed (n/a) option occurred on 95 occasions out of a possible 14,580 representing 0.65% of item scores. The overall missing data and not assessed rate was 0.85%.

7.3.3 Factor analysis

The 20 items of the APP for the total data set (n=729) were subjected to Principal Components Analysis (PCA) using SPSS version 14 (SPSS Inc.). Prior to performing PCA the suitability of data for factor analysis was assessed. Inspection of the correlation matrix revealed the presence of many coefficients of 0.3 and above. The Kaiser-Meyer-Okline value was 0.975, exceeding the recommended value of 0.6 and the Bartlett's Test of Sphericity reached statistical significance ($p=.000$), supporting the factorability of the correlation matrix. PCA demonstrated the presence of 1 dominant factor with an eigenvalue exceeding 1, explaining 59% of the variance as shown in Table 7.2.

Table 7.2: Component Matrix Field Test One

Component	Total Variance Explained			Extraction Sums of Squared Loadings		
	Initial Eigenvalues			Total	% of Variance	Cumulative %
	Total	% of Variance	Cumulative %			
1	11.758	58.791	58.791	11.758	58.791	58.791
2	.920	4.599	63.390			
3	.728	3.641	67.031			
4	.664	3.320	70.351			
5	.591	2.955	73.305			
6	.546	2.728	76.033			
7	.510	2.549	78.582			
8	.491	2.456	81.038			
9	.451	2.254	83.292			
10	.398	1.990	85.282			
11	.396	1.980	87.262			
12	.370	1.852	89.115			
13	.356	1.779	90.893			
14	.342	1.712	92.605			
15	.289	1.443	94.048			
16	.274	1.368	95.417			
17	.260	1.301	96.718			
18	.233	1.165	97.883			
19	.224	1.120	99.003			
20	.199	.997	100.000			

Extraction Method: Principal Component Analysis.

An inspection of the scree plot revealed a clear break after the first component (Figure 7.1).

Using the scree test, it was decided to retain only one component for further investigation.

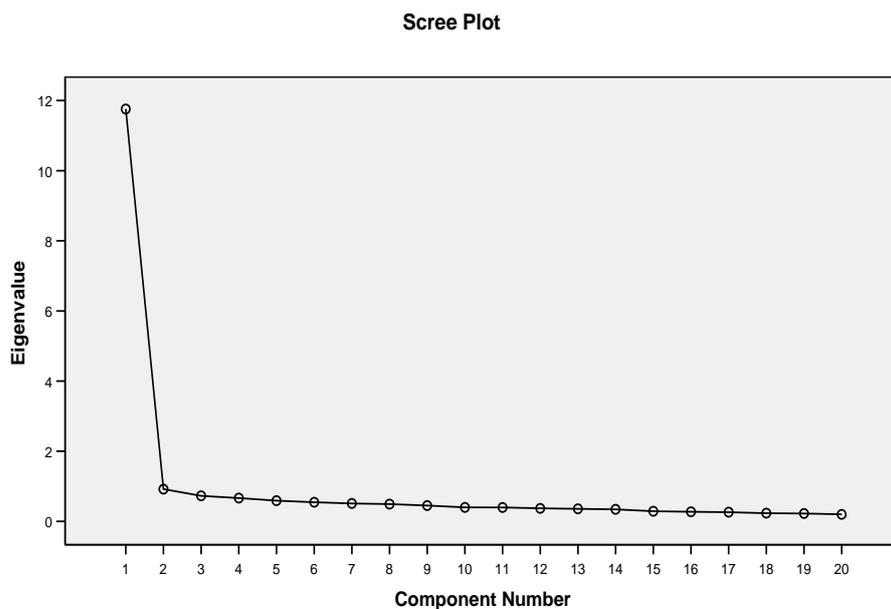


Figure 7.1: Scree plot

Retention of one factor was further supported by the results of the parallel analysis that showed only one component with an eigenvalue exceeding the corresponding criterion values for a randomly generated data matrix of the same size (20 variables x 729 respondents) (J. F. Pallant & Tennant, 2007). Parallel analysis is an additional method to determine the number of factors to retain and involves comparing the size of the eigenvalues with those obtained from a randomly generated data set of the same size (Horn, 1965). For this procedure the program developed by Watkins (2000), Monte Carlo PCA for parallel analysis was used. If the actual eigenvalue obtained from the PCA is larger than the criterion value from the parallel analysis then the factor is retained. This is demonstrated in Table 7.3. The results from parallel analysis support the decision from the screeplot to retain one factor.

Table 7.3: Factor analysis parallel analysis Field Test One

Component no.	Actual eigenvalue From PCA	Criterion value from Parallel analysis	Decision
1	11.75	1.30	Accept
2	0.92	1.24	reject
3	0.72	1.20	reject
4	0.66	1.17	reject
5	0.59	1.14	reject

The component matrix (Table 7.4) shows the loadings of each item on the first component. SPSS uses the Kaiser criterion (retain all eigenvalues above 1) as the default. Table 7.4 shows that all items load quite strongly (above 0.4) on the first component. This further supports the decision to retain only one factor. As only one component was extracted the solution cannot be rotated.

Table 7.4: Component matrix

	Component
	1
APP rating	.846
APP rating	.839
APP rating	.837
APP rating	.824
APP rating	.811
APP rating	.809
APP rating	.794
APP rating	.782
APP rating	.775
APP rating	.770
APP rating	.765
APP rating	.758
APP rating	.757
APP rating	.753
APP rating	.748
APP rating	.738
APP rating	.708
APP rating	.698
APP rating	.640
APP rating	.638

7.3.4 Rasch analysis: Overall Model Fit

Sample 1 (n=390)

The Chi-Square Item-Trait Interaction statistic was 111.8 ($df= 80, p= 0.01$) with the Bonferroni adjusted alpha (α) value = .0025 (.05/20). The chi-square probability value of $p = 0.01$ indicated fit between the data and the model.

Validation sample 2 (n=340)

The Chi-Square Item-Trait Interaction statistic was 88.63 ($df= 80, p= 0.23$) with the Bonferroni adjusted alpha (α) value = .0025 (.05/20). The chi-square probability value of $p = 0.23$ indicated fit between the data and the model.

7.3.5 Overall Item and Person Fit

Sample 1 (n=390)

The mean (SD) fit residual values for all items was -0.38 (1.82) indicating presence of some misfitting items to the model. Misfit of items indicates some deviation from the probabilistic

relationship between the individual item and the rest of the items on the scale. The residual mean value for persons was -0.36 (*SD* 1.38) indicating no misfit among the respondents in the sample.

Validation sample 2 (n=340)

The residual mean value for items was -0.24 (*SD* 1.81), again indicating some misfit of items to the model. Similarly the residual mean value for persons was -0.35 (*SD* 1.35) indicating no misfit among the respondents in the sample.

7.3.6 Individual Item and Person Fit

Sample 1 (n=390): Items

Two of the 20 items, item 6 (communication – written) and item 19 (applies evidence based practice to patient care), exhibited a positive item fit residual above 2.5 suggesting low levels of discrimination (Table 7.5). None of the items exhibited a significant chi-square value. Items 11 (Identifies and prioritises patient/client's problems) and 13 (selects appropriate intervention in collaboration with patient/client) displayed high negative fit residuals (-3.15 and -3.72) respectively (Table 7.5). This suggests some redundancy or over-discrimination in these items (Andrich, 1988).

Sample 1 (n=390): Persons

Examination of individual person-fit revealed four participants with positive fit residuals above +2.5. Investigation of these individual results revealed four instances of unexpected scoring on item 19 (evidence based practice), and one on item 3 (demonstrates ethical, legal and culturally sensitive practice). Deletion of this data made no difference to overall model fit.

Validation sample 2 (n=340): Items

Item 6 and 19 were again the only two items that exhibited a positive item fit residual above 2.5. None of the items exhibited a significant chi-square value. Again items 11 (Identifies and prioritises patient/client's problems) and 14 (performs interventions appropriately) displayed high negative fit residuals (-3.12 and -3.06) respectively (Table 7.5).

Validation sample 2 (n=340): Persons

There were four participants with positive fit residuals above +2.5. Investigation of these individual results revealed two instances of unexpected scoring on item 3 (demonstrates ethical, legal and culturally sensitive practice) and two instances of unexpected scoring on item 19. Again, deletion of this data made no difference to overall model fit.

Table 7.5: Individual item fit of 20 APP items to the Rasch model: Sample 1 (n=390) and sample 2 (n=340)
 (Item order is from least to most difficult of the 20 items)

Sample 1 (n=390)								Sample 2 (n=340)							
APP item	Location	SE	FitResid	DF	ChiSq	DF	Prob	APP item	Location	SE	FitResid	DF	ChiSq	DF	Prob
1	-1.79	0.111	0.481	366.46	7.056	4	0.132	1	-2.014	0.118	-0.05	318.02	5.89	4	0.207
3	-1.692	0.111	1.578	366.46	10.716	4	0.029	3	-1.893	0.117	1.697	318.02	9.113	4	0.058
2	-1.116	0.099	0.783	366.46	6.859	4	0.143	6	-1.505	0.116	2.845	318.02	5.861	4	0.209
5	-0.945	0.102	1.198	366.46	6.528	4	0.163	4	-1.028	0.105	1.543	318.02	8.996	4	0.061
4	-0.915	0.1	0.511	366.46	8.818	4	0.065	5	-1.026	0.109	1.188	318.02	6.545	4	0.161
6	-0.911	0.107	3.037	366.46	12.59	4	0.013	2	-0.593	0.106	1.323	318.02	5.066	4	0.280
7	-0.753	0.108	-0.176	366.46	5.15	4	0.272	7	-0.119	0.114	0.821	318.02	0.79	4	0.939
9	-0.445	0.11	-1.69	366.46	3.359	4	0.499	8	-0.099	0.106	-1.836	317.08	2.7	4	0.609
8	-0.324	0.105	-0.858	365.52	1.615	4	0.806	14	0.01	0.11	-3.064	318.02	5.977	4	0.200
20	-0.072	0.1	1.864	364.58	4.592	4	0.331	20	0.04	0.105	0.489	315.21	1.986	4	0.738
14	0.148	0.105	-2.494	366.46	11.602	4	0.020	9	0.13	0.118	-2.063	318.02	2.895	4	0.575
16	0.467	0.099	-0.895	365.52	0.923	4	0.921	18	0.398	0.109	-2.448	317.08	6.185	4	0.185
18	0.725	0.1	-0.673	354.24	1.007	4	0.908	11	0.411	0.109	-3.124	317.08	4.503	4	0.342
15	0.76	0.095	-0.428	366.46	0.792	4	0.939	13	0.723	0.112	-0.729	317.08	0.678	4	0.953
17	0.974	0.098	-1.584	366.46	1.118	4	0.891	15	0.784	0.101	-0.776	318.02	2.194	4	0.700
11	0.981	0.101	-3.153	363.64	4.549	4	0.336	16	0.965	0.106	-1.487	317.08	3.845	4	0.427
13	1.024	0.101	-3.72	365.52	7.583	4	0.108	10	1.039	0.108	0.104	307.7	2.199	4	0.699
12	1.124	0.102	-1.494	365.52	5.58	4	0.232	12	1.133	0.115	-0.392	317.08	2.048	4	0.726
10	1.189	0.099	-2.551	365.52	8.142	4	0.086	17	1.216	0.105	-1.688	318.02	1.351	4	0.852
19	1.57	0.097	2.852	358.94	3.307	4	0.507	19	1.428	0.101	2.971	312.39	9.818	4	0.043

Note: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies EBP 20 = assesses risk

7.3.7 Threshold ordering of polytomous items

There were no disordered thresholds for any of the 20 items in either sample one or two. The threshold map for sample one is illustrated in Figure 7.2. An additional example of the ordering of thresholds is illustrated in Figure 7.3 in the category probability curves for item 15 (is an effective educator) in sample 2.

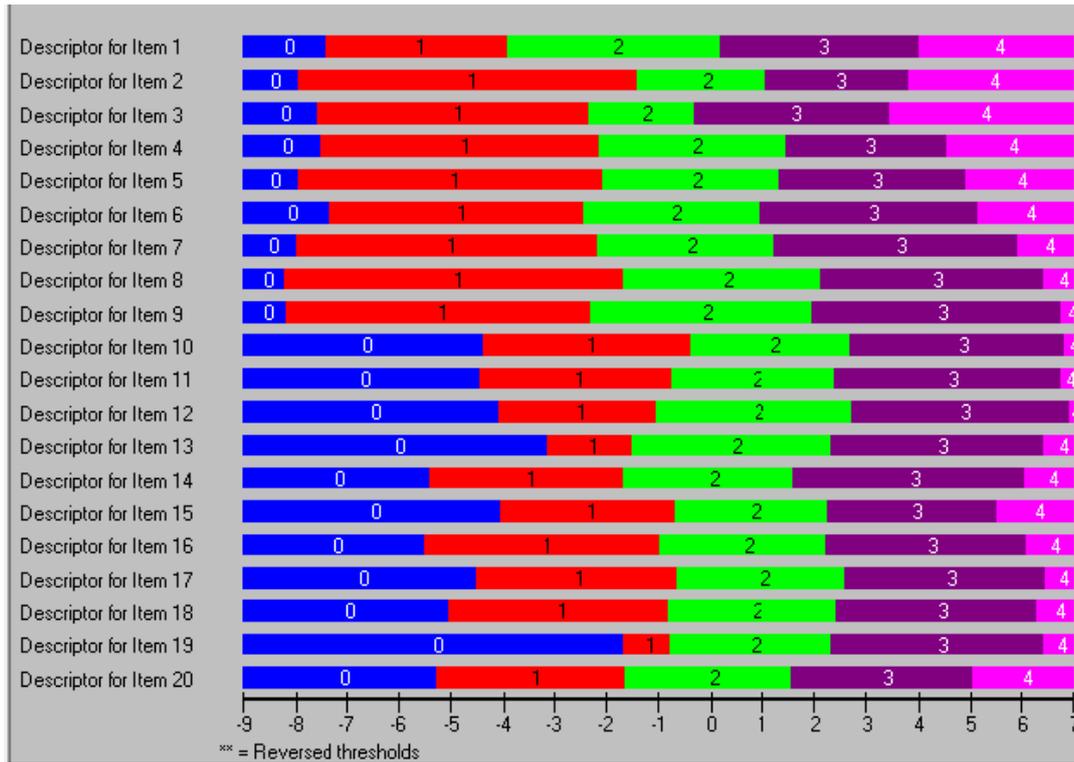


Figure 7.2: Threshold map of APP 20 items in sample 1 n=390

Legend: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies evidence based practice 20 = assesses risk

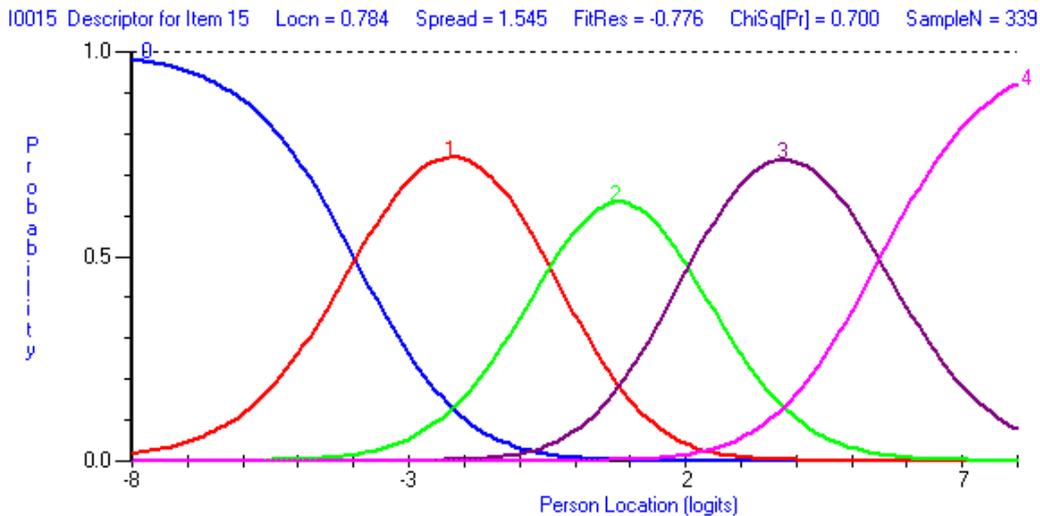


Figure 7.3: Category probability curves for item 15 (Is an effective educator) in data from sample 2 (n=340). Loqn=location; FitRes= Fit Residual; ChiSq[Pr]=chi-square probability

7.3.8 Targeting

Sample 1 (n=390) and Sample 2 (n=340).

Comparison of the mean fit residual scores obtained for the persons in each sample (-0.36 sample 1 and -0.32 sample 2) with zero set for the items provides an indication of how well targeted the items are for people in the current sample. For a well targeted measure the mean for the persons would also be around zero. The values of -0.36 and -0.32 would therefore indicate the instrument was well targeted for use with this group of students (J. F. Pallant, 2010).

Figures 7.4 and 7.5 allow visual inspection of the person item threshold graphs. The distributions of the students (top half of the graph) and item thresholds (bottom half of the graph) for the APP total score are presented on a logit scale for each sample. Visual inspection of these person item threshold graphs show that a majority of item thresholds correspond to the main cluster of persons (students). Logits of increasing negative value indicate less difficult items and less able students. Logits of increasing positive value indicate more difficult items and more able students.

Overall there was an even spread of items across the full range of student scores, suggesting effective targeting of the APP items. There were a few students at the extremes that were not covered. These may represent extremely high performing students or may be outliers

with particularly unusual scores. There are almost no students who are performing at a level too low to be captured by the scale.

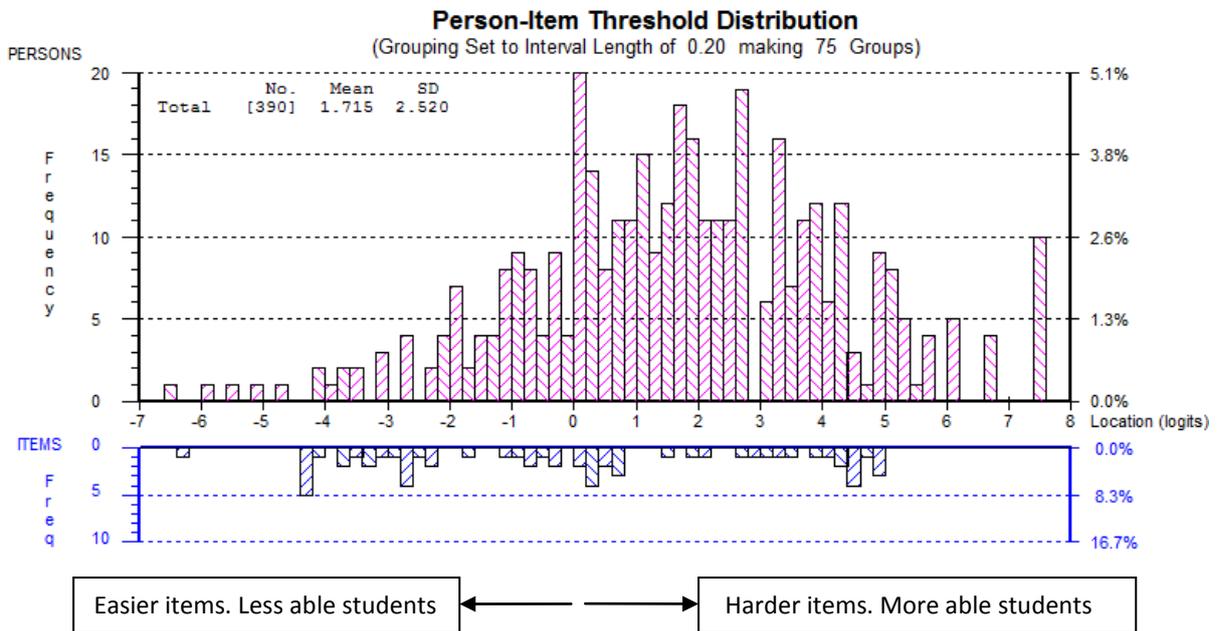


Figure 7.4: Person-item threshold distribution graph for sample 1 (n=390).

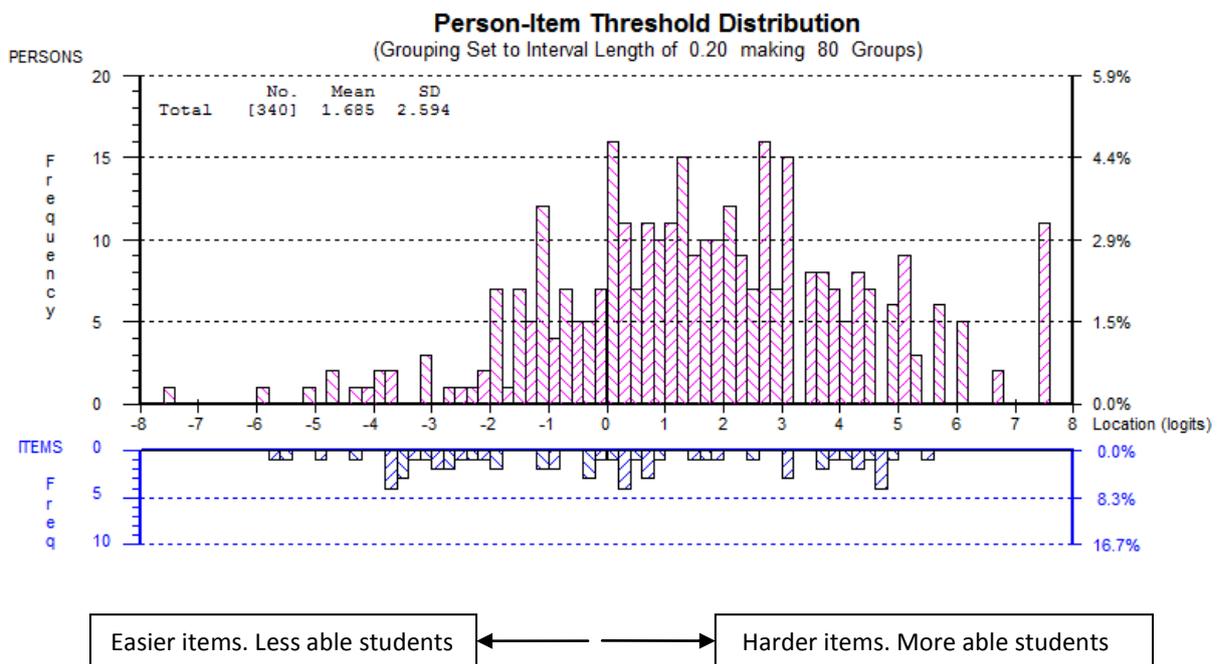


Figure 7.5: Person-item threshold distribution graph for sample 2 (n=340).

7.3.9 Hierarchy of item difficulty

The sequence or hierarchy of average difficulty of the 20 items on the APP for both samples are presented in Table 7.5 and graphically in Figure 7.6. In both samples the first six items representing professional behaviour and communication were the least difficult items whereas the most difficult items related to the application of evidence based practice to patient care and items relating to analysis and planning.

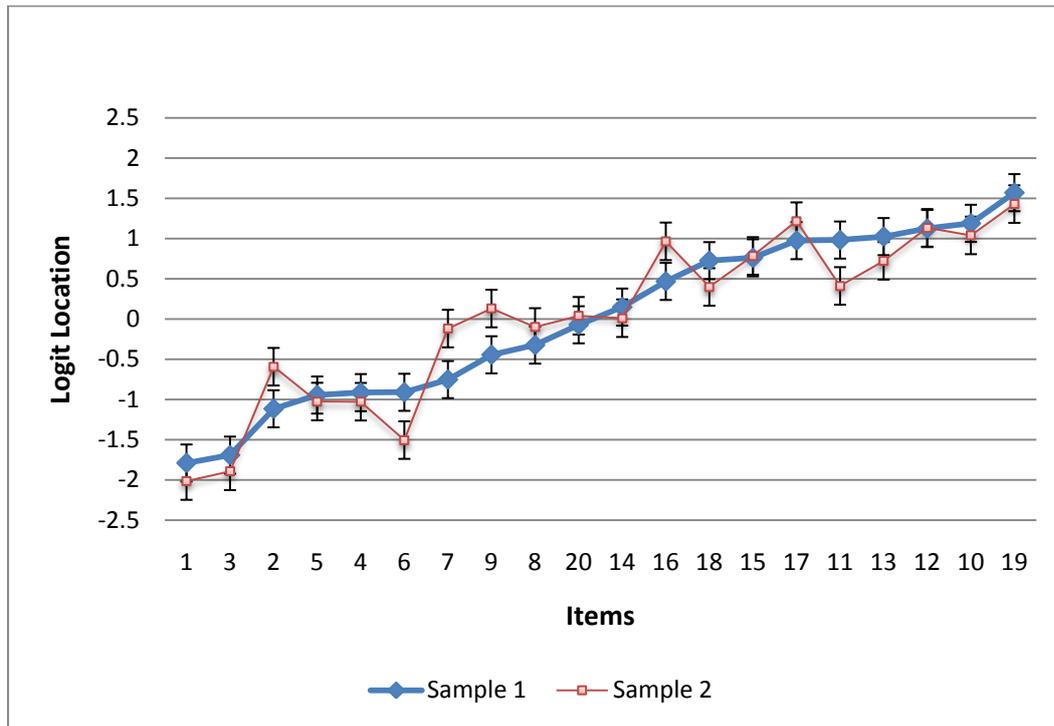


Figure 7.6: Logit location of APP items in two samples (sample 1 n=390; sample 2 n=340)

Note: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies Evidence Based Practice 20 = assesses risk

Figures 7.7 and 7.8 show the relationship between raw ordinal APP scores and person location logit scores for sample 1 and 2. (For complete data refer to Appendix 7.1)

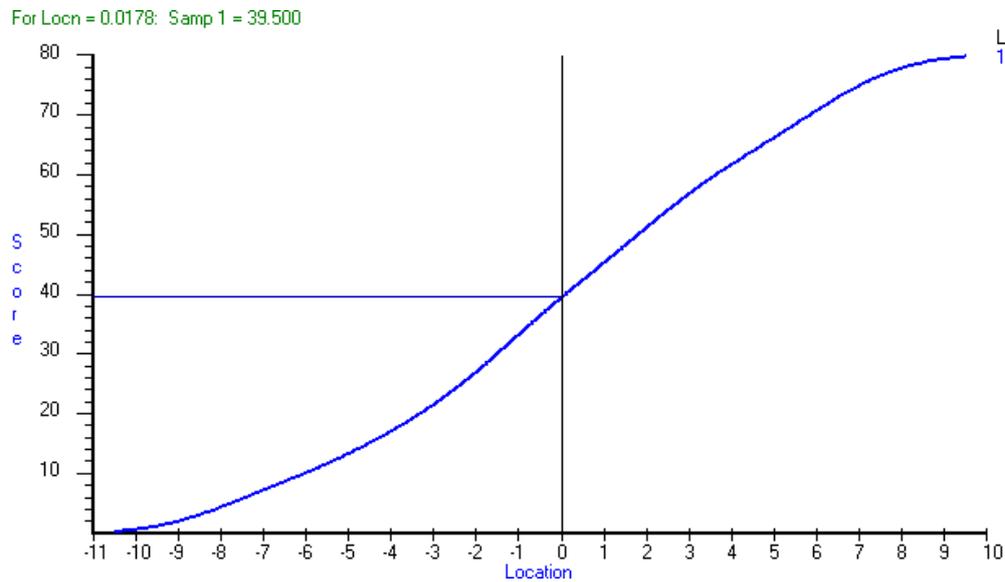


Figure 7.7: Plot of person logit location and raw APP score (sample 1)

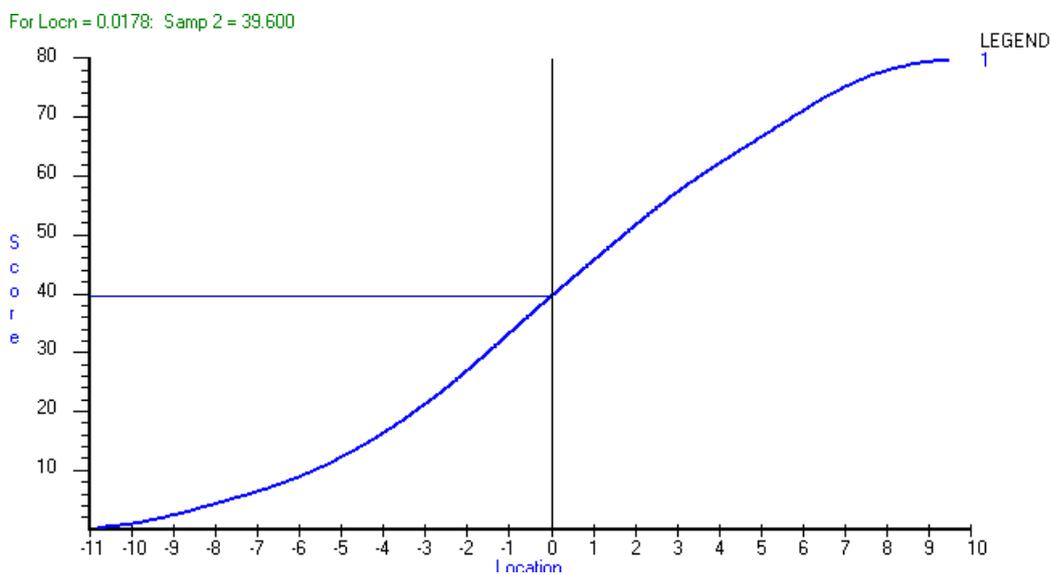


Figure 7.8: Plot of person logit location and raw APP score (sample 2)

7.3.10 Person separation index

Sample 1 (n=390) and validation sample 2 (n=340)

In sample one and two the PSI was 0.96 indicating the ability to discriminate between at least four levels of performance.

7.3.11 Differential item functioning (DIF)

The presence of item bias was explored by analysis of DIF with a Bonferroni-adjusted p value of .0025 (.05/20). No significant DIF was demonstrated in either of the two samples for the following variables: student age, clinical educator age, gender and experience as an educator, University, facility type and clinical area. APP item ratings were not systematically affected by any of these seven variables.

Item 6 (communication – written) demonstrated uniform DIF in sample 1 (n=390) with respect to student gender. In this sample the DIF for gender reaches statistical significance, .0008, (Table 7.6) which is less than the Bonferoni adjusted expected p value of greater than .0025. The presence of DIF for student gender did not occur in sample 2 (n=340) (Bonferonni adjusted $p = .006$).

Table 7.6: Uniform and non-uniform DIF statistics for all APP items for student gender for sample 1 (n=390)

Item	Uniform DIF				Non-uniform DIF			
	MS	F	DF	Prob	MS	F	DF	Prob
1	0.92357	0.90075	1	0.342891	0.73953	0.64052	4	0.634002
2	0.86886	0.86698	1	0.352104	0.79539	0.65615	4	0.622969
3	10.42662	7.78535	1	0.005399	2.15167	1.785	4	0.131845
4	0.04981	0.04516	1	0.831735	1.31372	1.2103	4	0.306533
5	0.26313	0.24705	1	0.619313	0.3854	0.38236	4	0.821194
6	14.20469	11.1262	1	0.000892	0.60939	0.43857	4	0.780702
7	1.68398	1.77544	1	0.183133	1.29849	1.50866	4	0.199679
8	0.40677	0.49727	1	0.480936	0.5802	0.65878	4	0.621124
9	1.60902	2.14336	1	0.143642	0.72016	0.99415	4	0.411021
10	0.92156	1.26456	1	0.261165	1.03033	1.38751	4	0.238294
11	0.04186	0.05862	1	0.808755	0.52328	0.60806	4	0.657138
12	0.19879	0.23025	1	0.631481	0.43137	0.48586	4	0.746127
13	0.03273	0.04667	1	0.828989	0.9597	1.42562	4	0.225497
14	4.91364	6.89004	1	0.008857	2.64323	3.4733	4	0.009623
15	2.23739	2.44731	1	0.118184	0.69844	0.92007	4	0.452553
16	0.31294	0.37122	1	0.542531	1.7112	1.85685	4	0.118104
17	1.2625	1.53148	1	0.216299	0.79854	0.95073	4	0.435023
18	0.48598	0.51646	1	0.472596	3.56048	3.66693	4	0.011266
19	4.18922	3.63405	1	0.057007	0.93098	0.86956	4	0.482558
20	0.18684	0.18236	1	0.66949	0.36155	0.40324	4	0.806264

Inspection of the item characteristic curve (Figure 7.9) suggests that at equal levels of the overall attribute, professional competence, male students are less likely than females to score highly on this item. Figure 7.10 shows the item characteristic curve for item 5 (verbal/non-verbal communication) for male and female students which provides an example of responses where there is no differential item functioning for student gender

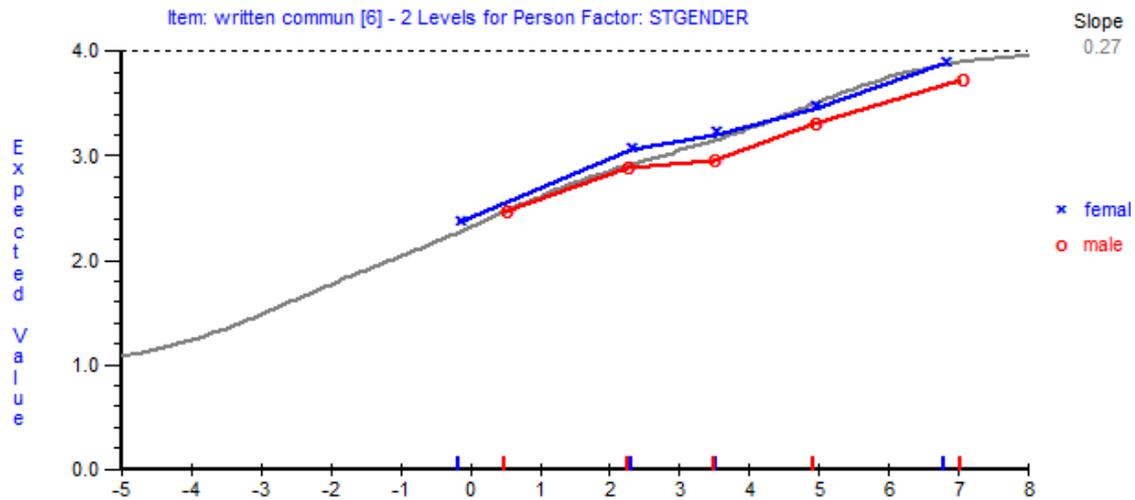


Figure 7.9: Differential item functioning graph of female and male students for item 6 (written communication) sample 1 (n=390)

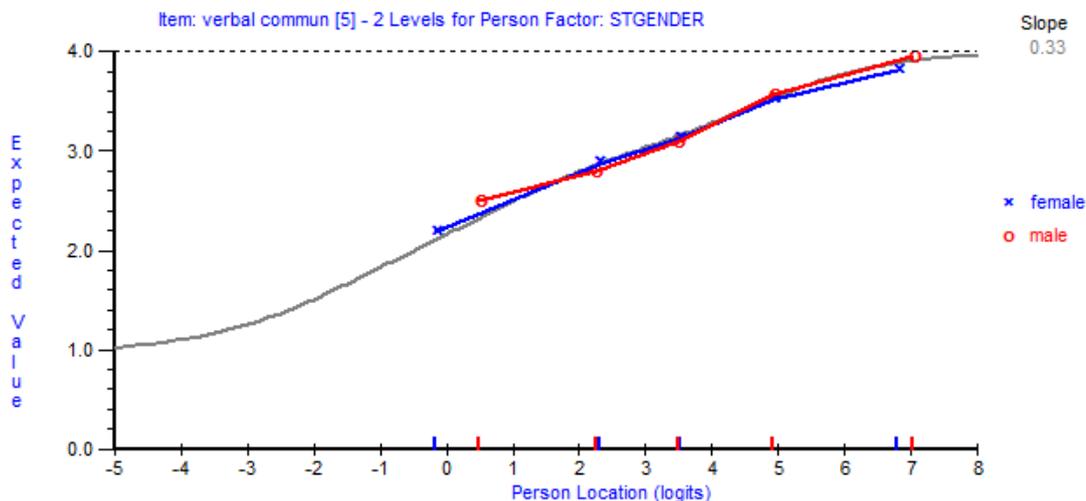


Figure 7.10: Differential item functioning graph of female and male students for item 5 (communication - verbal/non-verbal) sample 1 (n=390)

Student experience:

The level of student experience on clinical placements was examined for DIF by checking for the presence of DIF based on the number of weeks of clinical placement the student had

attended prior to Field Test One. The level of student experience was coded as beginning (0-9 weeks prior experience), middle (10-19 weeks prior experience) and end (20-35 weeks prior experience). The individual item scores and the global rating scale (GRS) for each completed APP were examined for DIF based on time. None of the items or the GRS showed probability values exceeding the adjusted alpha value (.0025) in either sample.

7.3.12 Dimensionality

Analysis of the pattern of item loadings on the first extracted factor of the residuals shows that the residuals loaded in opposite directions on two subsets defined by positive and negative loadings on the first factor. Only those items with loadings greater or less than 0.3 were considered. Local dependence occurs where the rating on one item influences how a later item is rated (Streiner & Norman, 2003). The items showing positive residual correlations in both samples were items 1 (demonstrates an understanding of patient rights and consent), 3 (demonstrates ethical, legal and culturally sensitive practice), 2 (demonstrates a commitment to learning) and 4 (demonstrates teamwork).

The next step was to investigate if the person estimates (location values) based on scores that underpin each of these sets of items were significantly different using independent t-tests. In sample one data 28 cases out of 390 (7.2% or 0.072) had statistically different scores on each of the subsets of items. A confidence interval for a binomial test of proportions was calculated for this observed number of significant tests. The 95% confidence intervals around this estimate are calculated as expected 19.5 ($=0.05 \times 390$), $(18.018 < 27.9981 \text{ (obs)} < 37.986)$ (Normal-z approx test) or as a proportion of 0.072, $0.046 < 0.072 \text{ (obs)} < 0.097$. As the expected ranges contained the observed value 19.5 or .05, unidimensionality of the scale is supported (Smith, 2002).

This analysis was repeated for sample 2 data. In sample two (n=340) data 24 cases out of 340 (6.82% or 0.068) or expected: 17.5 ($=0.05 \times 340$), $14.756 < 24.0006 < 33.252$ (Normal-z approx) or as a proportion of 0.068, $0.041 < 0.068 \text{ (obs)} < 0.094$. As the expected range contains the observed value 17.5 or .05, unidimensionality of the scale is supported.

7.4 Discussion

Data from the first field test of the APP instrument were overall consistent with the expectations of the unrestricted (partial credit) derivation of the Rasch model of measurement. This supports use of the 20 item APP as an instrument for measuring professional competence of physiotherapy students in the clinical environment.

Examination of the raw data revealed low levels of missing data. The level of missing data in this field test (.2%) is to that of the pilot trial (.3%).

Despite the overall low level of missing data, Item 19 (application of evidence based practice to patient care) was the item most often not scored or scored as not assessed (4.4%). Additionally, investigation of person fit residuals identified six instances of unexpected scoring on this item. This is a similar pattern to the pilot trial.

During the focus groups following both the pilot trial and Field Test One, clinical educators suggested that assessment of item 19 in the clinical context was often misunderstood (Chapter Five, section 5.4 and Chapter Six section 6.5.5). Consensus from these focus groups for more information in the clinical educator resource manual clarifying how to assess the application of evidence based practice (EBP) in patient care (item 19) supports the likelihood of educator confusion in relation to scoring of this item. In part this may relate to new approaches to undergraduate training, with recent graduates being trained in assessment of evidence and selection of practices that are aligned with best evidence. Graduates of previous years may not be as familiar with concepts taught to current students. It is likely that there are both differences in knowledge of EBP (comparing students to educators) and lower educator confidence in being able to grade students given limited training in research literacy for many clinical educators. As discussed in Chapter Five, these results also reflect those of earlier studies by Cross and Hicks (1997) and Cross et al (2001) where behaviours such as 'demonstrating research knowledge' proved to be difficult for educators to observe and assess.

While the data demonstrated overall fit to the Rasch model for both participant samples, item fit residual values indicated the presence of some mis-fitting items to the model.

Individual item fit statistics revealed problems with item 6 (written communication) and item 19 (application of evidence based practice to patient care) in both samples. Pallant and Tennant (2007) state that one of the most common sources of item misfit concerns respondents' (educators) inconsistent use of the scoring options resulting in what is termed 'disordered thresholds'. Investigation of threshold ordering of the 20 polytomous items showed there were no disordered thresholds in either sample implying that the cause of item misfit is not related to use of the rating scale by the educators. Scrutiny of differential item functioning however revealed the presence of uniform DIF for item 6 based on student gender in sample 1. Male students scored lower on item 6 than female students, in all five categories of professional competence from lowest to highest. In sample two there was no DIF for item 6 based on student gender. Item 19 did not show any evidence of DIF for student gender in either sample. While the presence of DIF for gender for item 6, written communication, is cause for some initial concern, several factors support retention of the item at this stage and review following the second field test. These factors include that the APP instrument is still in the early stages of development and refinement, the validation sample showed no signs of DIF for gender and the importance of the item to the overall professional competence skill set.

No significant DIF was demonstrated in either of the two samples for the variables student age and experience, clinical educator gender, age and experience as an educator, University, or clinical area. The absence of DIF for these variables is a critical component of research into inconsistencies related to assessment of student performance in work-place settings. All stakeholders can be confident that APP ratings were not systematically affected by any of these variables.

It has been suggested by previous authors that student scores can be expected to increase across time, that is, as the level of student experience on clinical placements increases there will be a concomitant increase in their overall scores (Task Force for the Development of Student Clinical Performance Instruments, 2002). Examination of the raw total scores showed a progressive increase in the mean total score on the 20 items and the mean score on the global rating scale as the number of weeks of student clinical experience increased. Rasch analysis however demonstrated that none of the items or the GRS demonstrated the

presence of DIF for level of student experience. The absence of DIF for level of student experience does not imply that there is no change in student scores across time, but rather that the instrument is ranking students in a similar way at different time points.

The targeting of the instrument shows adequate coverage of thresholds across the whole construct of professional competence. The gap between item-thresholds on the logit scale observed in pilot data (Chapter Four section 4.3.8) was not observed in this field trial. There were only a few students with higher ability than even the most difficult items. Given that both samples contained students completing clinical placements just prior to graduation, it is reasonable to expect that there will be some high performing students.

The sequence or hierarchy of average difficulty of the 20 competencies on the APP provides an indication of which clinical competencies may be easier to acquire eg communication and professional behaviours and those that are more difficult and therefore may be expected to take longer to master, e.g. application of evidence based practice, analysis and planning (critical thinking), goal setting and progression of interventions. Focus group discussions following Field Test One revealed that the hierarchies of both samples also fit closely with the experience of clinical educators regarding items they observe to be more difficult for students to master (refer to Chapter Six). Focus group discussion however, did not support the finding of item 19 being the most difficult item for students to master. Rather, educators reported difficulty in knowing how to assess this item appropriately in the clinical environment. The confusion educators reported in how to assess this item may be the reason underpinning the position of item 19 as the hardest item to master and the fact that the item showed misfit to the model. Further investigation of item 19 in the second field test may provide insight into the functioning of this item.

The validation of item hierarchy on the second sample analysis indicates the robust nature of the observed item difficulties and supports consistency in item scores across samples.

In a previous study using Rasch analysis to investigate the functioning of a clinical assessment instrument for physiotherapy students, Rheault and Coulson (1991) demonstrated the ranking of items from easiest to hardest to be similar to the results of this study. From easiest to most difficult the items were: exhibits professionalism, exhibits communication skills, performs effective treatment skills, performs safe treatment skills, can

problem solve and works from an adequate knowledge base. While comparison of results from Rasch analysis using different software packages is problematic, the similarity with the hierarchy of average difficulty in this study is evident.

Prior research on the Clinical Performance Instrument (CPI) using the Rasch measurement model (WINSTEPS program), demonstrated that clinical instructors were able to discriminate among six levels of performance (Straube & Campbell, 2003). The CPI scoring system consisted of a 100 mm VAS (101 point scale) and an additional 'with distinction' box. This scoring system could be considered equivalent to a scale with 102 points or categories. Straube and Campbell (2003) demonstrated that only when the CPI scoring system was collapsed to 6 categories did the scale function with distinct scoring groups, that is, as a valid rating scale. In this research, pilot trial and Field Test One data demonstrated that the five level scale exhibited high reliability (PSI=.96) with no disordered thresholds, indicating a highly discriminative scale. This means that educators were able to identify four levels of student ability using the five point scale on the APP. Hence educators can be confident in the ability of APP scores to identify more and less able students.

The Rasch measurement model is a unidimensional model, therefore the assumption is that the items form a unidimensional scale. To test this assumption a principal components analysis of the residuals allows for a test of the local independence of the items (Smith, 2002; Wright, 1996b). The APP includes items relating to communication, professional behaviour and physiotherapy specific skills but nevertheless appears robust when tested against the assumptions of the Rasch measurement model, with the independent t-test analysis supporting the assumption of unidimensionality.

In order to comprehensively investigate the functioning of the APP, factor analysis was also performed. While not considered necessary by Rasch analysts, factor analysis is a traditional component of statistical analysis of instrument scores applied by classical test theorists and provides additional evidence for scale unidimensionality. Additionally, as discussed in Chapter One, other instruments used to assess professional competence in physiotherapy students have been investigated using factor analysis demonstrating them to be internally reliable multidimensional instruments, measuring at least two constructs, physiotherapy specific clinical skills and professional behaviour. Factor analysis of Field Test One data

determined the presence of one dominant factor explaining 59% of the variance supporting unidimensionality of the APP. The fact that earlier research found two factors and the data presented here demonstrated only one dominant factor is not readily explained, but could be speculated to be related to difference in samples or statistics used.

One of the primary advantages of Rasch analysis is that raw ordinal scores may be converted to interval level Rasch scores. Conversions from raw scores to Rasch scores can be provided, but this adds a layer of complexity to calculating the student's final score that it appears can be avoided due to the almost perfect linear relationship shown in Figures 7.7 and 7.8. These Figures demonstrate a linear relationship, with slight flattening at scale extremes which indicates that raw scores can be used with confidence as though they were interval, unless scores are at the extremes.

7.5 Actions arising following Field Test One

Following consideration of Field Test One qualitative data (Chapter Six) and quantitative data in this Chapter, a number of changes were made to the APP (version 3). A summary of the modifications is presented in Table 7.7. The amended APP instrument (version 4) for use in Field Test Two is provided in Appendix 7.2.

Table 7.7: Modifications to APP (version 3) following Field Test One

Requested modifications to APP (v3)	APP (version 3) used in field test 1	APP (version 4) used in field test 20
<p>Change wording on the global rating scale (GRS). Replace poor with not adequate and satisfactory with adequate to align with language used in the rating scale</p> <p>Modify wording on item 13 as educators considered wording placed too much emphasis on patient collaboration rather than on selection of intervention</p> <p>Item 11: add in additional PI to highlight the importance of clinical reasoning</p>	<p>GRS: In your opinion as a clinical educator, the overall performance of this student in the clinical unit was: Poor; Adequate; Good; Excellent</p> <p>Item 13: Collaborates with patient/client to select appropriate intervention</p> <p>Identifies and prioritises patient's/client's problems</p> <ul style="list-style-type: none"> • generates a list of problems from the assessment • collaborates with the patient/client to prioritise the problems • considers patient's/clients values, priorities and needs 	<p>GRS: In your opinion as a clinical educator, the overall performance of this student in the clinical unit was: <i>Not Adequate; Adequate; Good; Excellent</i></p> <p><i>Item 13: Selects appropriate intervention in collaboration with patient/client.</i></p> <p>Identifies and prioritises patient's/client's problems</p> <ul style="list-style-type: none"> • <i>generates a list of problems from the assessment</i> • <i>collaborates with the patient/client to prioritise the problems</i> • <i>considers patient's/clients values, priorities and needs</i> • <i>justifies prioritisation of problem list</i> • <p>Selects and measures relevant health indicators and outcomes</p> <ul style="list-style-type: none"> • <i>selects all appropriate variable/s to be measured at baseline from WHO ICF domains of impairment, activity limitation and participation restriction.</i> • <i>identifies and justifies variables to be measured to monitor treatment response and outcome.</i> • <i>selects appropriate tests/outcome measures of each variable for the purpose of diagnosis, monitoring and outcome evaluation.</i> • <i>links outcome variables with treatment goals</i> • <i>communicates the treatment evaluation process and outcomes to the client</i> • <i>identifies, documents and acts on factors that may compromise treatment outcomes</i>
<p>Modify wording on PIs for item 8 to ensure comprehensive, clear, and are more aligned with WHO ICF domains.</p>	<p>Selects appropriate methods for measurement of relevant health indicators</p> <ul style="list-style-type: none"> • <i>selects important, functional and meaningful outcomes relevant to treatment goals, including those to identify potential problems</i> • <i>chooses appropriate methods/instruments to measure identified outcomes across relevant assessment domains e.g. impairment, activity limitations, participation restriction, well-being and satisfaction with care</i> 	<p>Selects and measures relevant health indicators and outcomes</p> <ul style="list-style-type: none"> • <i>selects all appropriate variable/s to be measured at baseline from WHO ICF domains of impairment, activity limitation and participation restriction.</i> • <i>identifies and justifies variables to be measured to monitor treatment response and outcome.</i> • <i>selects appropriate tests/outcome measures of each variable for the purpose of diagnosis, monitoring and outcome evaluation.</i> • <i>links outcome variables with treatment goals</i> • <i>communicates the treatment evaluation process and outcomes to the client</i> • <i>identifies, documents and acts on factors that may compromise treatment outcomes</i>

Legend: WHO: world health organization; ICF: International classification of function; GRS: global rating scale, (Code: amendments to instrument highlighted in *italics*)

7.6 Chapter Summary

Following on from the results of the pilot trial, this chapter has presented the results of Field Test One providing further data on the measurement properties of the APP as investigated by Rasch analysis and forms a platform for discussion and ongoing refinement of the APP instrument.

8. Chapter Eight: Field Test Two - Qualitative evaluation.

8.1 Introduction

In response to the results of the first field test several modifications were made to the APP instrument and training manual (Table 7.7). Given these changes a second field test was required to address outstanding questions generated during the first field test, to evaluate the refinements made to the APP instrument and clinical educator training manual and to test the validity of Field Test One results. Similar to the first field test, qualitative and quantitative data on APP instrument performance were gathered to assist in assessing the validity of the measurements and feasibility of its use in the practice environment. Ideally the refinements made to the APP instrument following the first field test would assist in standardising use of the instrument and enhance the overall performance of the instrument. The unresolved issue of the most appropriate number of scoring categories and number of items in each domain warranted continued evaluation. Finally investigation of items 6 (written communication) and 19 (applies evidence based practice), both quantitatively and qualitatively, was of particular importance.

This chapter describes the three stages of Field Test Two: Stage 1: Preparation for Field Test Two, Stage 2: During the field test and Stage 3: On completion of testing. In addition the chapter presents the results of analyses of qualitative data collected during these stages. Chapter Nine then focuses on the results of Rasch analysis of Field Test Two quantitative data and summarises the final modifications made to the APP instrument following completion of the second and final field test.

8.2 Methods

8.2.1 Stage 1: Preparation for Field Test Two

8.2.1.1 Modifications to APP resource manual and training workshops

In line with the iterative action research design of the study illustrated in Figure 2.5, following analysis of, and reflection on, the quantitative and qualitative results from the first field test several modifications to the training manual and workshops for educators and students were made prior to the second field test. The changes were made by the author

(MDal) if the qualitative data demonstrated consensus for the change and the change did not run counter to the results of quantitative data analysis (Chapter Six section 6.5). Modifications to the resource manual included additional information in the frequently asked questions section clarifying how to assess time management and application of evidence based practice. In addition a section in the resource manual was included outlining potential educator beliefs and behaviours related to assessment that are best avoided (refer to Appendix 8.1, p 16). While not a specific recommendation from the focus groups, this information was included to assist educators to standardise their approach to assessment and was based on comments made by educators and students during focus groups. The training workshops were modified to include the following: specific examples on how to assess written communication (item 6) and evidence based practice (item 19), clarification of the changes made to the wording of items 8, 11, 13 and the GRS and how to appropriately assess them, correct use of the rating category not applicable (N/A), and how to use the instrument to provide effective mid unit formative feedback.

8.2.1.2 Modifications to feedback questionnaires

Based on data collected during the first field test (Chapter Six), modifications were made to the educator and student feedback questionnaires prior to Field Test Two.

Clinical educator feedback questionnaire (Appendix 8.2)

Feedback from educators following the first field test indicated face to face training was the preferred method for training. However, given this is not always feasible, it was important to investigate what other training methods were acceptable to clinical educators. To address this, the following question was added to the clinical educator feedback questionnaire.

Given face to face training in the use of the APP is not always possible, please indicate which of the following training options you would find effective.

- Teleconference
- Self directed learning package (includes manual and CD/DVD)
- On-line training module (ie training module completed on-line)
- Other (please specify).....

Student feedback questionnaire (Appendix 8.3)

Data from the student feedback questionnaire completed during Field Test One showed that students considered the training they received on assessment during clinical practice and use of the APP was comprehensive. Despite this result, there was feedback from the students during Field Test One focus groups that the level of engagement with the written material provided in the student's university clinical education policy and procedure manuals was inconsistent. In addition, involvement of the students in reflection on their performance prior to mid unit formative feedback and signing off on and receipt of a copy of the final completed APP instrument was also variable in Field Test One. In order to clarify these issues, six dichotomous questions were added to the student questionnaire for completion during Field Test Two:

1. I read the section on assessment in the policy and procedures manual
2. I attended the lecture about clinical assessment given at university prior to commencing my clinical unit
3. I received a copy of my mid unit feedback comments from my educator
4. I received a copy of the completed APP at the end of the unit
5. I signed off on my end of unit assessment results
6. I self scored on the APP prior to mid unit feedback meeting with my educator

Respondents were asked to rate their agreement to 10 statements using a 5 point scale (1 = Strongly Disagree and 5= Strongly Agree). Open ended questions were included to gather comprehensive feedback.

8.2.1.3 Demographic data form

As in Field Test One a demographic data form was completed by each clinical educator and student participating in the second field test. Since investigation of the same factors that might bias measurements (differential item function) was to be conducted during Field Test Two, the same demographic form used in Field Test One was used in the second field test.

8.2.1.4 Recruitment of participants

To maintain consistency, the same nine universities who participated in the first field test were approached and agreed to participate in Field Test Two and the same method used in Field Test One was employed in Field Test Two. Information on the research (provided in face to face meetings and in writing) was provided to students and clinical educators undertaking major clinical placements from each participating university (Appendix 9.1). Following presentation of information on the second field test interested students and educators were invited to sign forms consenting to the analysis of their deidentified assessment and questionnaire data. Assessment data were excluded from analysis if either the student or their clinical educator did not consent to participation in the research. As in Field Test One, of the nine programs participating in Field Test One, six programs agreed to use the APP in parallel with their current university specific clinical assessment form. Three programs elected to use the APP as the sole assessment instrument (Table 6.1).

8.2.1.5 Training of participants

Compulsory workshop attendance for all clinical educators participating in Field Test Two was not feasible in the authentic clinical education environment. Repeating the methods used in Field Test One, all clinical educators received training through workshop attendance and/or access to the APP resource manual. Students were educated in the assessment process and use of the APP instrument by a member of the research group or the clinical education manager at each university prior to commencing their clinical unit (Chapter Six, section 6.2).

8.2.2 Stage 2: during Field Test Two

To maintain consistency the same method as used in Field Test One was used in the second field test. Prior to commencement of the clinical unit each participating clinical educator received a field test package containing the resource manual, a copy of the APP (version 4) instrument for each student, a clinical educator and student demographic data form and feedback questionnaire. A reply paid envelope was provided to facilitate return of completed forms.

Students from the nine participating universities completed clinical units ranging in length from four to six weeks with all students engaged full time in clinical education. The clinical units represented the major areas of physiotherapy practice and included musculoskeletal, cardiorespiratory, neurological, paediatric and gerontological physiotherapy. For the six programs using the APP in parallel with a current university-specific form, the educators were instructed to assess the student's performance using the APP at the end of the clinical unit prior to completing the required university assessment documents. In these six programs, students did not view the completed APP instrument. Mid unit formative feedback was provided using the current university-specific form. In the three university programs where the APP was the sole assessment instrument, the educators completed the APP at mid and end of the unit. During end of unit summative assessment, students viewed the completed APP instrument.

Similar to Field Test One, teleconferences for educators were conducted mid way during each clinical unit. All teleconferences were conducted by one member of the research group (MDal) and attendance was not compulsory but recommended. The teleconferences provided support for the clinical educators and assisted in standardising information about the use of the APP. Any issues raised during teleconferences were recorded in an APP issues register and collated at the end of Field Test One using the coding guide presented in Table 8.5. This enabled the research group to determine the issues that concerned the educators and the frequency and occurrence of consensus of participants regarding each issue. The flow chart presented in Figure 6.1 summarises Field Test Two procedures with the exception that no 'think aloud' interviews were conducted during the second field test.

8.2.3 Stage 3: on completion of Field Test Two

8.2.3.1 Data management and analysis

On completion of each placement completed forms were returned by mail to the project manager (MDal) and entered into Microsoft Excel 2003. All data were de-identified once entered into spread sheets for statistical analysis and names of educators, students and physiotherapy programs were replaced by codes. Data were checked for accuracy and the links between names and codes were then permanently destroyed. Similar to the first field

test data analyses for Field Test Two were performed using SPSS 14.0 (SPSS Inc.) and RUMM2020 software (Andrich, et al., 2003) for Rasch analysis (refer to Chapter Seven).

8.2.3.2 Focus groups conducted following Field Test Two

Participants

As in previous focus groups conducted following the pilot (Chapter Five, section 5.2.1) and first field test (Chapter Six section 6.4.1) educators and students participating in the second field test were invited to attend a focus group. Participants were screened to enable representativeness of the larger stakeholder population. Recruitment was designed to optimise representation of all stakeholders by location (metropolitan, regional/rural and remote), clinical area of practice, years of experience as a clinical educator/supervisor or manager, organization (private, public, hospital based, community based and non-government). Participation in focus groups was voluntary and all participants consented prior to attendance at a focus group.

Duration and site selection

Each focus group was scheduled for one and a half hours and arranged at a time and location to suit participants.

Moderators

As with previous focus groups the principal moderator for each focus group was a research assistant with expertise in focus group methods. The research assistant was not a physiotherapist and had no vested interest in the outcomes of the research, reducing the likelihood of moderator bias.

Questions

Each group commenced with the moderator reminding the group that anything said during the group interview should remain confidential, that the session would be audio taped and that the transcription would be de-identified. Questions introduced in the educator focus groups covered the items and performance indicators (content, wording and clarity), scale (size, format, pass level, understanding of levels of performance), layout of the instrument,

and training in the use of the APP. The student group questions focussed on their experience of the instrument and how it was used during the clinical unit (Chapter Six section 6.4.2). As saturation of data during the focus group was achieved the questions were adapted and broadened to ensure comprehensive data collection.

On completion of Field Test Two qualitative data from all sources (training workshops, teleconferences, focus groups, and questionnaires) were analysed individually and then collated and sorted by topic and content. Analysis of both the qualitative (Chapter Eight) and quantitative (Chapter Nine) data provided the basis for refinement of the APP after the second and final field test.

8.5 Results: Qualitative evaluation Field Test Two.

Ethics approval was obtained from the Human Ethics Committee of Griffith and Monash Universities and from the Human Ethics Committees of each university where a physiotherapy program leader had agreed to participate in data collection in either the pilot trial or any of the subsequent field tests (n = 2) (Appendix 3.4).

8.5.1 Participant characteristics

In Field Test Two, a total of 663 completed APP assessments from 456 students were returned by 298 clinical educators. Nineteen of the 663 (2.8 %) completed assessments were deemed incomplete leaving 644 completed assessments available for analysis. Nine university physiotherapy programs participated in the field test. The 456 students were completing clinical placements of 4 – 6 weeks duration during the last 18 months of their physiotherapy program. Table 8.1 provides a summary of participant characteristics.

Table 8.1: Field Test Two: participant characteristics (n=644)

9 Universities: Australia=7, NZ=2	Characteristics
Student (n=456)	
Age (years) <i>Mean ± SD</i>	23.0 ± 3.5
Age range (years)	20 - 48
Gender %	66% F
Missing data	12%
CE (n=355)	
Age (years) <i>Mean ± SD</i>	34.1 ± 8.3
Age range (years)	22 - 60
Gender	72%F
Years of experience as CE <i>Mean ± SD</i>	6.4 ± 5.56
Range (years of experience)	0 – 34
% CEs attending training as an educator (prior to field test training)	56% yes, 16% no, (28% missing)
Self rating of experience as a clinical educator <i>Mean ± SD</i> (1= no experience – 5= very experienced)	3.46 ± 1.02
Missing data	16%
Time taken to complete APP (mins) <i>Mean ± SD</i>,	29.04 ± 19.3
Range (mins)	8 - 120
Clinical area (% of time spent in clinical area during unit)	
Cardiorespiratory physiotherapy	23.0
Neurological physiotherapy	25.0
Musculoskeletal physiotherapy	32.0
Paediatric physiotherapy	6.0
Speciality units e.g., spinal injuries, burns, women's health, oncology, mental health, hand therapy, plastic surgery	5.5
Missing data	8.5
Patient/client age group (%)	
Children (0-12 years)	4.0
Adolescents(13-20 years)	3.5
Adults (21-65 years)	51.5
Older persons (> 65 years)	36.0
Missing data	5.0
Type of facility (%) n=423	
Public hospital	54
Private hospital	7
Community based services	9
Private practice	3
Non-government organisation	6
Missing data	21
University Program (% of completed assessments n=644)	
La Trobe	19.0
Monash	28.0
Griffith	20.0
JCU	7.5
Uni Syd	6.0
Curtin	6.7
Otago	1.0
CSU	3.3
AUT	1.0
Missing data	7.5

8.5.2 Participant training: Workshops, and teleconferences

A total of 234 clinical educators attended workshops held across Australia and New Zealand as part of training for clinical educators involved in Field Test Two. All students received information regarding the research and were prepared for each clinical unit by a member of the research group or by a university clinical education manager. In addition, during the clinical units, 57 educators attended 10 teleconferences (Table 8.2).

Table 8.2: Field Test Two clinical educator training

Locations	No. of participants
Victoria	
Angliss Hospital, Box Hill Hospital, Maroondah Hospital, The Alfred Hospital, Caulfield General Medical Centre, Northern Health Network, Monash University Peninsula Campus and Gippsland Campus.	55
Western Australia	
Curtin University, Charles Gairdner Hospital	24
New South Wales	
Charles Sturt University, Northern Rivers Health District (Lismore, Ballina, Port Macquarie, Grafton), Newcastle University.	48
Queensland	
<u>Brisbane</u> and surrounding districts Toowoomba Hospital, Queen Elizabeth II Hospital, The University of Queensland, Griffith University (Gold Coast Campus)	63
<u>Far North Queensland</u> : Townsville Hospital including videoconference to surrounding districts (Cairns, Mackay, Proserpine, Mt Isa), Rockhampton Hospital	18
New Zealand	
Otago University, Auckland University of Technology	26
Australia and New Zealand	
Mid unit clinical educator Teleconferences (n=9)	57

8.5.3 Feedback questionnaires

Two hundred and twenty-two (74.5%) clinical educators and 251 (55%) students returned a feedback questionnaire. Some feedback sheets were incomplete. Questionnaires were returned by educators and students from all nine universities, by student across year levels, and a representative spectrum of placement and facility types.

8.5.3.1 Clinical educator feedback questionnaire

The results of the questionnaire are presented in Table 8.3.

Table 8.3: Clinical educator feedback on APP (n=222)

Question	Mean \pm SD*	Median
Confident using 0 – 4 rating scale	4.0 \pm 0.7	4.0
Confident using Global Rating Scale	4.1 \pm 0.8	4.0
APP practical in the clinical environment	4.3 \pm 0.7	4.0
Performance Indicators (PIs) useful	4.2 \pm 0.6	4.0
PIs easy to understand	4.2 \pm 0.6	4.0
Time taken to complete APP acceptable	4.2 \pm 0.6	4.0
Beginning practitioner definition helpful	4.1 \pm 0.7	4.0
Scoring rules helpful	4.2 \pm 0.7	4.0
Resource manual information on how to complete the APP was comprehensive	4.3 \pm 0.6	4.0
Preference for on-line version	4.0 \pm 0.7	3.0

*Each item rated from 1 = strongly disagree to 5 = strongly agree.

Completing the APP on-line

In response to the question “In the future, I would prefer to complete the APP on-line rather than posting/faxing hard copies”, forty-one percent of clinical educators agreed or strongly agreed that completing the APP on-line was preferable while twenty-eight percent disagreed, preferring to continue posting in hard copies. Thirty-one percent were ambivalent.

Training preferences

In response to the additional question; 'Given face to face training in the use of the APP is not always possible, please indicate which of the following training options you would find effective?' The results were as follows (Table 8.4):

Table 8.4: Preferred training options

Training Options	Yes
Teleconference	20%
Self directed learning package (includes manual and CD/DVD)	20%
On-line training module (ie training module completed on-line)	43%
Other (indicating that any of these methods, in isolation or in combination were acceptable)	5%
Missing data	12%

8.5.3.2 Student feedback questionnaire

The results of the questionnaire are presented in Tables 8.5 and 8.6.

Table 8.5: Student feedback on APP (n=251)

Question	Mean*	Median
Confident CE used 0-4 scale correctly	3.7± 1.1	4.0
Pls useful to assess own performance	4.0± 0.8	4.0
Scoring rules appropriate	4.0± 0.7	4.0
Entry level performance (pass) was clear to me	4.1± 0.9	4.0
Items easy to understand	4.2± 0.6	4.0
APP practical for use in clinical environment	4.0± 0.6	4.0
Performance required to score 4 was clear to me	3.8± 1.3	4.0
Information about APP prior to unit was adequate	3.7± 0.9	4.0
Rating on GRS was a fair indication of my performance	4.0± 1.1	3.0
Rating on 20 items were a fair indication of my performance	3.7± 1.2	3.0

*Each item rated from 1 = strongly disagree to 5 = strongly agree.

Table 8.6: Student feedback questionnaire results – dichotomous questions

Student feedback questionnaire (n=243)	Yes%	No%
I read the section on assessment in the policy and procedures manual	76	24
I attended the lecture about clinical assessment given at University prior to commencing my clinical unit	94	6
I received a copy of my mid unit feedback comments from my educator	86	14
I received a copy of the completed APP at the end of the unit	84	16
I signed off on my end of unit assessment results	90	10
I self scored on the APP prior to mid unit feedback meeting with my educator	46	54

The four open ended questions on the student questionnaire provided information matching that obtained in the first field test (Chapter Six section 6.5.4). Since no new information was obtained, data saturation was considered to have been reached.

8.5.4 Focus groups conducted following Field Test Two

One focus group for clinical educators (n=9), and two groups for students (n=15) were conducted on completion of Field Test Two. Application of a coding guide to the focus group data enabled identification of themes and assisted interpretation of the data. The coding guide (Table 8.7) was similar to that used in analysis of previous focus group data. (Chapter Three, section 3.3 and Chapter Six, section 6.4.1). An additional theme, acceptability, was added to the coding guide to enable collation of specific data concerning this area.

Table 8.7: Coding guide for content analysis of focus groups

Code	Content
It	Items: content, wording, clarity of intent, weighting of items
Sc/GRS	Scale: size, format, wording, sensitivity, global rating scale
Pass	Pass standard: passing performance,
PIs	Performance Indicators: perceived utility, number, content, clarity of intent, wording, suggestions for additional PIs
IFor	Instrument format: layout of instrument, perceived utility, suggestions for improvement
Tr	Training in the use of the APP: requirements of a training package,
Fback	Use of the instrument in providing feedback
Accept	Acceptability/utility of instrument use in the clinical context, time taken to complete
Other	Other key words, ideas, themes

8.5.4.1 Clinical educators

Content analysis of the second field test focus group data demonstrated findings comparable to Field Test One, suggesting saturation of the data had been reached. A summary of the clinical educator focus group data is presented in Table 8.8.

Acceptability of the APP

The majority of educators agreed that use of a common assessment instrument by all Universities eased the burden relating to assessment of student performance and enabled standardisation of training. Both aspects were influential in the consensus of participants on acceptability of the instrument. The following educator quote summarises this theme;

“Standardized assessments throughout all clinical schools would be great as we assess students from four different universities here and the paperwork and training can be overwhelming.”

Training

Training workshops were well received and considered essential to maintain standardisation in use of the instrument within and across facilities, clinical areas and educators of differing levels of experience. Video exemplars demonstrating passing and excellent performances

were suggested as a strategy to assist both students and educators to understand and use the scoring system. Production of these video exemplars had commenced.

Global rating scale (GRS)

Focus group data showed that use of the GRS by educators was more consistent than in the previous field test. There was consensus from participants regarding the wording changes on the GRS from poor, satisfactory, good and excellent to inadequate, adequate, good and excellent. Participants agreed that alignment of the wording on the GRS with that of the scoring definitions facilitated more consistent use of entry level as the passing benchmark during summative assessment, rather than the norm referenced approach used by some educators in Field Test One. Similar to Field Test One, use of the GRS by educators to provide mid-unit formative feedback remained inconsistent.

Item 6 - Written communication

Focus group discussion revealed division between educators on the acceptable minimum standard for written communication. The discrepancy in minimum standard occurred primarily in relation to writing up patient/client clinical notes rather than discharge summaries and letter writing to referring doctors or referral to other health professionals. Some educators considered the taking of notes during the patient interview to be essential, while others felt note taking impacted on effective communication with the patient and required students to write up notes on completion of the interview or during a scheduled break. Where there were facility based clinical assessment proformas, students were generally expected to complete these during the patient interview. The clinical area also influenced educator's expectations. In the acute inpatient hospital ward setting, the expectation was to write up multiple chart entries after assessing and treating several patients. The availability of patient charts in a busy ward environment was cited as the main reason underpinning this approach. In the outpatient setting, patient notes were often completed during the patient interview. Overall the expectations of educators varied and were often not discussed with the student prior to commencement of the clinical placement.

In summary, the instrument was viewed as user friendly and time efficient while maintaining comprehensive coverage of domains of clinical practice requiring assessment by a majority of educators. This theme is reflected in the following educator comments;

“This instrument is very time efficient compared with current assessment forms in use.”

and

“I initially found it somewhat time consuming to fill out APP at mid unit as the form was comprehensive. However, by end of unit it was very efficient as I had a foundation to build on.”

Table 8.8: Summary clinical educator focus group results Field Test Two

Target issue and discussion	Outcomes/ Decisions/ Actions arising	Group decision
Items	<ul style="list-style-type: none"> • Number of participants reporting difficulty assessing item 19 reduced since Field Test One. Plan to monitor this item in future field tests. • Item 6 identified as an issue. Participants from different clinical areas, divided on what is acceptable minimum standard. 	Monitor item 19
<ul style="list-style-type: none"> • Difficulty assessing item 19 evidence based practice experienced by some educators in Field Test Two • Item 6 (written communication). 		No consensus on item 6 rating
Scale size	<ul style="list-style-type: none"> • Rasch analysis of pilot trial & Field Test One results showed scale was working appropriately, therefore agreement that five level scale to remain with two failing levels and two passing levels. • Consideration of having both a five level scale for some items and a pass/fail scale for some items. Review following field test. 	No consensus on scale size.
<ul style="list-style-type: none"> • <i>Number of scoring categories</i>: discussion concerning increasing number of scoring categories or maintaining 5 categories. • <i>Pass/fail categories</i>: discussion regarding whether some items could be graded pass/fail only, eg., risk management. 	<ul style="list-style-type: none"> • Global rating scale very useful addition to the instrument. • Modification of wording had improved the GRS. More consistent use by educators of entry level as the benchmark for adequate. • Wording changed: n/a replaced with not assessed 	Consensus to continue with current scale.
Scale		Consensus
<ul style="list-style-type: none"> • Global rating scale • Not assessed category: some educators reported that n/a could be interpreted as not applicable rather than not assessed and requested a change to the wording. 		
Performance indicators	<ul style="list-style-type: none"> • Performance indicators are comprehensive and helpful for provision of feedback and for rating of student performance 	Consensus
<ul style="list-style-type: none"> • <i>Use of performance indicators in assessment</i>: very positive feedback that PIs are helpful to guide mid unit feedback and assist end of unit summative grading. • performance indicators provide clear and constructive feedback. 		
Acceptability of APP	<ul style="list-style-type: none"> • Time taken to complete APP is acceptable • APP is very user friendly and comprehensively covers all aspects of practice requiring assessment 	Consensus
Training	<ul style="list-style-type: none"> • Training package/resource manual meets clinical educator needs (comprehensive, practical and accessible). • Training well received and needs to continue once research complete. Video exemplars of students demonstrating passing and excellent performances were requested as a strategy to assist in training of both students and educators 	Consensus on outcomes
<ul style="list-style-type: none"> • <i>Standardisation</i>: Use of one instrument by all Universities will allow for standardisation of training. • <i>Training format</i>: preference is for face to face training but on-line training module would ensure accessibility by all educators • <i>Information in training package</i>: content is comprehensive 		
Other issues	<ul style="list-style-type: none"> • Future development of on-line version of APP • Development of on-line self directed module 	Consensus
<ul style="list-style-type: none"> • Participants agreed that a single instrument used by all Universities was an important step for physiotherapy profession. • Development of on-line version of APP, web-based discussion board 		

8.5.4.2 Students

Content analysis of the second field test student focus group data also demonstrated similar findings to Field Test One. The student responses were again more personally focussed with less homogeneity in responses than was found in the clinical educator groups (Table 8.9).

Table 8.9: Summary student focus group results Field Test Two

Target issue	Student focus group results
Items and domains of practice	<ul style="list-style-type: none"> • All students agreed item content was comprehensive • Item 19, evidence based practice, was not well understood by clinical educators • Often confused about exactly what item 18 ‘undertakes discharge planning’ actually meant, ie, how to demonstrate competence in this area.
Rating scale	<ul style="list-style-type: none"> • Half of the student participants considered need for additional scoring categories i.e., six or seven categories rather than five • Occasional comments that different educators (in the one unit) had different ideas on performance required to score a 2, 3 or 4.
Feedback	<ul style="list-style-type: none"> • Often verbal feedback and scoring on items did not correlate. Verbal feedback positive but scoring low. • Some educators scored items as ‘1’ at mid unit to make student work harder in the second half of the clinical unit • If more than one educator, feedback could vary between educators • Performance in the first week of the unit often influenced your final unit score. • Overall consensus that educators could provide more regular feedback
Performance Indicators (PIs)	<ul style="list-style-type: none"> • All students considered PIs very useful to guide assessment particularly during mid unit formative feedback • All students considered PIs were comprehensive and easy to understand
Overall, I consider the scores I received for each of the 20 items were a fair indication of my performance	<ul style="list-style-type: none"> • Wide variety in responses from students, with some agreeing and others considering their marks were too low. No student considered their scores were too high.
Training for students	<ul style="list-style-type: none"> • Majority of students were satisfied with training provided on APP and assessment during clinical unit, although several students admitted to non-attendance at training sessions • Presentation and information provided in clinical manuals very comprehensive. • Video exemplars of students demonstrating passing and excellent performances were requested as a strategy to assist in training of both students and educators

8.6 Discussion

Similar to Field Test One, qualitative data from the second field test, demonstrated strong convergence in the opinions of educators and students regarding several aspects of the APP instrument including, comprehensive content coverage of items and performance indicators, ease of use of the instrument within the clinical context, importance of performance indicators in providing clear, well-targeted feedback on performance, role of training of all stakeholders in assessment processes, and the effectiveness of the resource manual in providing accessible information on assessment practices. Analysis of qualitative data also highlighted several unresolved issues requiring further investigation. In view of the similarity of qualitative data from both field tests, the discussion in this chapter will present any differences and additional issues that arose in the qualitative data collated following Field Test Two.

In regard to the items and performance indicators, item 19 (application of EBP) appeared to be better understood and assessment more consistent across educators. Focus group data suggested that additional information included in the resource manual assisted in educators applying a more consistent approach to assessment of this item. Data from focus groups, feedback questionnaires and training workshops showed that scoring of item 6 (written communication) was varied, with the minimum expected standard of performance inconsistently interpreted by educators from different clinical areas. Feedback obtained during the training workshops conducted during Field Test Two would suggest that educators had no insight into these different expectations, considering instead that all educators had similar viewpoints to themselves on how to assess written communication. From a medico-legal perspective standardisation of the minimum standard of written communication required is essential. This issue warrants further investigation and training of educators.

Modifications to the wording on the global rating scale (GRS) were well received by educators. There appeared to be more consistent use of entry level standard as the benchmark for student performance to be rated at an adequate level. There remained inconsistent use of the global rating scale at mid unit with some educators still scoring the

GRS in relation to the student's prior experience rather than against entry level performance as instructed. The persistence of this behaviour when providing mid unit formative feedback led the research team to amend the APP instrument, advising educators that the GRS was to be completed only at end of unit summative assessment.

Educators again reported that the APP was acceptable for use in the clinical environment. Despite the mean time to complete increasing from 17 minutes in the first field test to 29 minutes, educators remain satisfied that this does not represent an unnecessary burden.

Feedback from student focus groups and feedback questionnaires demonstrated that further clarification of behaviours required to achieve a rating of four for an item was required. Students requested video exemplars of students demonstrating passing and excellent performances were requested as a strategy to assist in training of both students and educators. Development of exemplars by the research group had commenced and would be included as part of the final version of the APP resource manual (Chapter Twelve).

Self reported attendance at training in use of the APP was high with 96% of students indicating they had attended some training at university prior to their clinical unit. While this level of attendance is to be admired, it may have been influenced by the research nature of the field testing. Future tracking of this statistic is warranted.

Seventy-five per cent of students reported reading the section on assessment in the policy and procedures manual resource manual. Comments from students indicated an understanding of the importance of being prepared for their clinical assessment, tinged with honesty in relation to whether or not they had actually read the information provided in the resource manual.

The majority of students reported receiving a copy of the mid unit feedback provided to them by their educators, and 90% reported signing off on their final summative assessment. Conversely the level of engagement of students in reflection on their performance at mid unit using the APP instrument was only 46%. This indicates that greater pre clinical unit training for students on the essential role reflection plays in promoting continuing improvement in performance and as a life-long skill of effective health professionals is required.

Students indicated a slight reduction in satisfaction levels when responding to the question “the ratings I received on the 20 items were a fair indication of my performance”. In Field Test One 68% either agreed or strongly agreed with the statement, while in Field Test Two 64% agreed or strongly agreed. As students become more educated in the process of assessment of clinical performance, their rejection of unsubstantiated ratings of items is likely to increase. The importance of timely and specific, feedback supported by evidence remains paramount. It is important to note that a potential level of bias in the findings relating to user satisfaction could also be present since only those clinical instructors and students who self selected to return their questionnaires were included in reported results.

Clear consensus was achieved during student and educator focus groups on the importance of training in assessment practices. Educators reported a preference for face to face training but were aware that this was not always achievable. The majority of educators agreed that on-line self directed learning modules were an acceptable alternative mode of training. Similarly preference for completion of the APP instrument on-line increased from 24% to 41% between the two field tests.

8.7 Chapter Summary

Overall, the data demonstrated strong convergence in the opinions of educators and students regarding several aspects of the APP instrument and also highlighted several unresolved issues requiring further investigation and continued management.

9. Chapter Nine: Field Test Two - Quantitative evaluation

9.1 Introduction

As highlighted in the previous Chapter, the aims of the second field test were to address outstanding questions generated during the first field test, to evaluate the refinements made to the APP instrument and clinical educator training manual and to test the validity of Field Test One results. Analyses of qualitative data from Field Test Two have been reported in Chapter Eight, while this Chapter focuses on the results of Rasch analysis of Field Test Two quantitative data and continues investigation of validity evidence supporting the proposed use of the APP as a measure of work-place based clinical performance of physiotherapy students.

9.2 Validity evidence based on relations to other variables

As described in Chapter One, the Standards for Educational and Psychological Testing (1999) advise that evidence of validity be assembled from multiple sources to substantiate the planned interpretations of instrument scores. One of the sources of evidence relates to the relationship between instrument scores and other variables. This source of evidence is primarily based on correlational studies and seeks to provide both confirmatory (convergent), and discriminant evidence of test score validity (Downing, 2003; Streiner & Norman, 2003). Another important purpose of this external aspect of validity is to document, where relevant, anticipated between-group and within-person changes over time concerning the target construct (Wolfe & Smith, 2007b).

9.2.1 Convergent and discriminant evidence

Evidence of relational validity of the APP instrument was investigated by examining the correlation between APP scores and other measures intended to assess either a construct related to clinical performance (convergent evidence) or an unrelated construct hypothesised not to be strongly correlated with workplace performance scores (discriminant evidence) (Messick, 1996; Wilson, 2005). One approach is to correlate data from an existing assessment instrument with the newly developed APP. However because there are no gold standard instruments for assessing professional competence in

physiotherapy students, correlating data from a newly developed instrument and an existing assessment instrument was considered not beneficial in developing an argument for the validity of the APP. An alternative approach was to compare the APP scores of a subset of students completing a clinical unit during Field Test Two with the results on three other independent university based assessment tasks to establish what relationship if any existed between the different assessment tasks.

9.2.2 Developmental progression in competency

Previous research on assessment of professional competence of physiotherapy students hypothesised and then demonstrated that with increasing time spent in clinical practice there is a significant increase in ratings of student performance (P. D. Cox, et al., 1999; Fitzgerald, et al., 2007; Loomis, 1985a; Task Force for the Development of Student Clinical Performance Instruments, 2002) (Chapter One). Each of these four studies used this data to provide support for the validity of their instrument scores. In the first field test the mean APP score was shown to increase as expected when students were divided into groups with increasing amounts of placement experience (Chapter Seven). In this field test the relationship of individual student (within-person) scores to hours of experience was examined to continue exploration of the hypothesis that as students are exposed to increasing hours of clinical experience, their scores on the APP instrument would reflect this.

9.3 Method

9.3.1 Field Test Two

The procedure for Field Test Two replicated that of Field Test One and has been described in Chapter Eight.

9.3.2 Validity evidence based on relations to other variables

9.3.2.1 Convergent and discriminant evidence

To investigate convergent and discriminant evidence of validity, the APP scores of a subset of students (n=94) from Griffith University completing an orthopaedic inpatient clinical unit

during Field Test Two were compared with the results on three other independent university based assessment tasks to establish what relationship if any existed between the different assessment tasks. This subset of students was a sample of convenience from the university where the author (MDal) worked, whose data could be readily de-identified and for which ethics permission was obtained. The three assessment tasks were a two hour written examination, a thirty minute practical skill-based examination, and a written assignment on orthopaedic radiology. While all three assessment items were focussed on testing the students' knowledge and skills in relation to orthopaedic physiotherapy, it was hypothesised that the results from the written examination and radiology assignment were testing at the levels of 'knows' and 'knows how' on Miller's (1990) pyramid and would correlate poorly, if at all, with the results of the orthopaedic practical exam and orthopaedic clinical unit which were hypothesised to be testing at the levels of 'shows how' and 'does'. Additionally it was hypothesised that the results of the orthopaedic practical exam and orthopaedic clinical unit would be correlated as they were designed to examine practically based skill sets of the students at the higher levels of 'shows how' and 'does'. The three University based assessment items were components of a Griffith University physiotherapy course completed by undergraduate physiotherapy students.

At the time of collecting the data, the research question now asked had not been developed, so it is unlikely that knowledge of research hypothesis could have influenced the relationship between grades for tasks. In addition as per university procedure, assessors for each assessment item were independent and blind to grades obtained by students for the other items. No grades for these pre-clinical course assessment items were available to the educators assessing students during their clinical unit. Additionally, student results on the three University based assessment tasks were from the previous semester and as such marks had already been ratified by the University and were not able to be altered.

Information on this aspect of the research was provided to the students and their consent was obtained. Students were advised that all data would be de-identified prior to analysis. As Griffith University was using the APP instrument in the second field test in parallel with the Griffith University specific clinical assessment form, students had been informed that the scores on the APP would not be used in establishing their grade for the clinical unit.

9.3.2.2 Developmental progression in competency

To investigate evidence of validity relating to within-person changes over time concerning the target construct, the APP scores for a subset of students (n=57) from Monash University were analysed. While all students attended one University there were no *a priori* reasons to consider that this group of students would be different from those attending other Australian universities and data from Field Test One had demonstrated that there was no differential item functioning for University attended by students. Hence a sample of convenience, available for review at the time this chapter was developed, was used for analysis. This subset represented all students who had completed six clinical units (3 of 4 weeks and 3 of 5 weeks duration) during the last 18 months of their undergraduate degree. The APP scores for each student across the six units were analysed to establish if there was a significant change in student scores as their time within the clinical environment increased, and when those changes occurred. Information on this aspect of the research was provided to the students and their consent was obtained. Students were advised that all data would be de-identified prior to analysis.

9.3.3 Data management and analysis

Using methods developed and tested in the pilot trial and Field Test One, completed student assessment forms were returned to one of the researchers (MDal) by mail, entered into a spreadsheet and de-identified. Data analyses were performed using SPSS 14.0 (SPSS Inc.) and RUMM2020 software (Andrich, et al., 2003) for Rasch analysis. Pearson product-moment correlations for four individual assessment tasks and a one-way repeated measures ANOVA comparing APP scores at six consecutive time periods were conducted to investigate validity evidence based on relations to other variables. Raw scores (0-4) for each of the twenty items on the Field Test Two version (version 4) of the APP (Appendix 6.5) and the sum of item scores (a total score) for each student assessed during the field test were examined using Rasch analysis. Similar to Field Test One exploratory factor analysis with parallel analysis was conducted prior to Rasch analysis.

9.4 Results

Ethics approval was obtained from the Human Ethics Committee of Griffith, La Trobe and Monash Universities and from the Human Ethics Committees of each university where a physiotherapy program leader had agreed to participate in data collection in either the pilot trial or any of the subsequent field tests (Appendix 3.4).

9.4.1 Participants' characteristics – Field Test Two

In Field Test Two a total of 644 APP completed assessments from 456 students were returned by 298 clinical educators. Nine university physiotherapy programs participated in the second field test. The 456 students were completing clinical placements of 4 – 6 weeks duration during the last 18 months of their physiotherapy program. Table 8.1 presents a summary of participant characteristics.

9.4.2 Characteristics of item and instrument scoring

Table 9.1 presents the descriptive statistics of the raw scores for each item, the raw total scores for the 20 summed item scores and the frequencies of use of each rating scale category for the 20 items on the 644 completed assessments.

The overall mean of the total APP scores for 644 completed assessments was 60.9 / 80 (76.1%) ($SD = 11.98$). The mean rating on the global rating scale for 644 APP forms was 3.1 where (1= inadequate, 2= adequate, 3= good, 4= excellent). For students with 0-10 weeks of clinical experience prior to Field Test Two, the mean APP score was 53.82/80 ($SD = 11.4$), for students with 10-20 weeks prior experience the APP mean score was 61.69/80 ($sd = 12.8$) and for 20-30 weeks, 65.34/80 ($SD = 11.2$). The mean (SD) APP score for students receiving a GRS of inadequate was 34.84 (6.39) for adequate 50.39 (4.00), for good 62.91 (4.11) and excellent 73.49 (4.05).

Table 9.1: Descriptive statistics Field Test Two (n=644)

Item	N valid	Mean	Standard Error of Mean	SD	Rating 0		Rating 1		Rating 2		Rating 3		Rating 4		N/A	
					Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
1	644	3.43	.027	.68	0	0	4	0.6	53	8.2	233	36.2	354	55.0	1	0.15
2	644	3.32	.031	.79	1	0.2	13	2.0	84	13.0	217	33.6	329	51.1	1	0.15
3	644	3.49	.027	.69	0	0	3	0.5	47	7.3	206	31.9	388	60.2	1	0.15
4	643	3.21	.033	.82	1	0.2	14	2.2	103	16.0	245	38.0	280	43.6	4	0.62
5	644	3.12	.030	.76	1	0.2	12	1.9	104	16.1	298	46.3	229	35.6	0	0
6	644	3.19	.029	.74	2	0.3	5	0.8	87	13.5	306	47.5	244	37.9	1	0.15
7	642	3.05	.028	.72	1	0.2	8	1.4	115	19.3	335	71.4	183	52.2	1	0.15
8	644	2.87	.029	.73	1	0.2	34	5.3	166	25.7	329	51.1	114	17.8	2	0.31
9	644	2.86	.028	.71	0	0	18	2.8	155	24.0	345	53.5	126	19.7	0	0
10	644	2.78	.030	.77	1	0.2	23	3.6	140	21.7	364	56.5	116	18.1	0	0
11	644	2.88	.032	.80	3	0.5	27	4.2	153	23.7	308	48.0	153	23.7	0	0
12	643	2.78	.031	.78	2	0.3	29	4.5	171	26.6	326	50.8	115	17.9	1	0.15
13	642	2.85	.031	.79	1	0.2	32	5.1	164	25.6	322	50.1	123	19.1	0	0
14	644	2.96	.030	.77	1	0.2	27	4.2	162	25.2	314	48.8	140	21.7	0	0
15	643	2.97	.033	.83	0	0	21	3.3	129	20.0	323	50.3	170	26.5	1	0.15
16	643	2.89	.031	.78	0	0	25	3.9	144	22.4	277	43.0	197	30.7	0	0
17	641	2.81	.031	.78	1	0.2	18	2.8	161	25.1	308	48.1	153	23.9	0	0
18	634	2.86	.032	.80	0	0	28	4.4	160	25.2	302	47.6	144	22.8	0	0
19	639	2.90	.032	.82	2	0.3	32	5.0	138	21.6	310	48.4	157	24.6	8	1.24
20	642	3.03	.033	.83	3	0.5	18	2.8	122	19.0	289	44.9	210	32.7	5	0.77
GRS	642	3.1	.029	0.76	Not applicable to GRS	Not Adequate		Adequate		Good		Excellent		Not applicable to GRS		
						16	2.5	104	16.2	306	47.6	217	33.7			
Tot. score for 20 items		60.99 /80	.047	11.98	Range of total raw scores for 20 items: minimum=16; maximum=80											

Legend: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies evidence based practice 20 = assesses risk; N/A= not assessed; SD.= standard deviation; N=number; GRS=global rating scale

The data presented in Table 9.1 show that the total score for the 20 items ranged from 16 to 80, and all 5 points on the 0 – 4 rating scale were used for the majority of items. Item 18 (undertakes discharge planning) and item 19 (applies evidence based practice in patient care) were the items most frequently not scored. The missing data rate for item 18 was (1.5%), item 19 (0.77%) and the overall missing data rate was 0.21% (28 items not scored out of a possible 12,880 item scores). The frequency of use of not assessed (n/a) option occurred on 26 occasions out of a possible 12,880 representing 0.21% of item scores. The overall missing data and not assessed rate was 0.41%.

9.4.3 Characteristics of orthopaedic examination results

Table 9.2 presents the descriptive statistics of the results for the three university based orthopaedic examinations and the APP scores for the 5 week orthopaedic clinical unit.

Table 9.2: Descriptive statistics of orthopaedic examination results (n=94)

Assessment Task /100	Range	Min	Max	Mean	Std. Error	SD	Variance
radiology assignment	38.0	49.0	87.0	66.50	.84	8.15	66.42
practical examination	48.5	50.0	98.5	77.08	1.19	11.56	133.72
written examination	33.0	55.5	88.5	74.85	.70	6.81	46.50
clinical unit (APP score)	51.4	47.4	98.8	79.90	1.13	11.04	121.88

9.4.3.1 Pearson's product-moment correlation coefficient

The results of Pearson product-moment correlation for each of the four assessment tasks is presented in Table 9.3.

Table 9.3: Correlations between different orthopaedic examination formats

		clinical unit	written exam	radiology exam	practical exam
clinical unit	Pearson Correlation	1	.104	.050	.311**
	Sig. (2-tailed)		.317	.632	.002
	N	94	94	94	94
written exam	Pearson Correlation	.104	1	.325**	.189
	Sig. (2-tailed)	.317		.001	.069
	N	94	94	94	94
radiology exam	Pearson Correlation	.050	.325**	1	-.087
	Sig. (2-tailed)	.632	.001		.403
	N	94	94	94	94
practical exam	Pearson Correlation	.311**	.189	-.087	1
	Sig. (2-tailed)	.002	.069	.403	
	N	94	94	94	94

There was a weak but significant positive correlation between the results of the clinical unit and the practical skill examination [$r = .31, n=94, p < .002$] (Cohen, 1988). There was no correlation between either of the written assessment tasks (written exam and radiology assignment) and the practically focussed assessment items. There was also a weak but significant positive correlation between the two written assessment items [$r = .33, n=94, p < .001$]. Visual examination of the scatter plot (Figure 9.1) of the clinical unit and practical skill exam demonstrates the widespread nature of the data points.

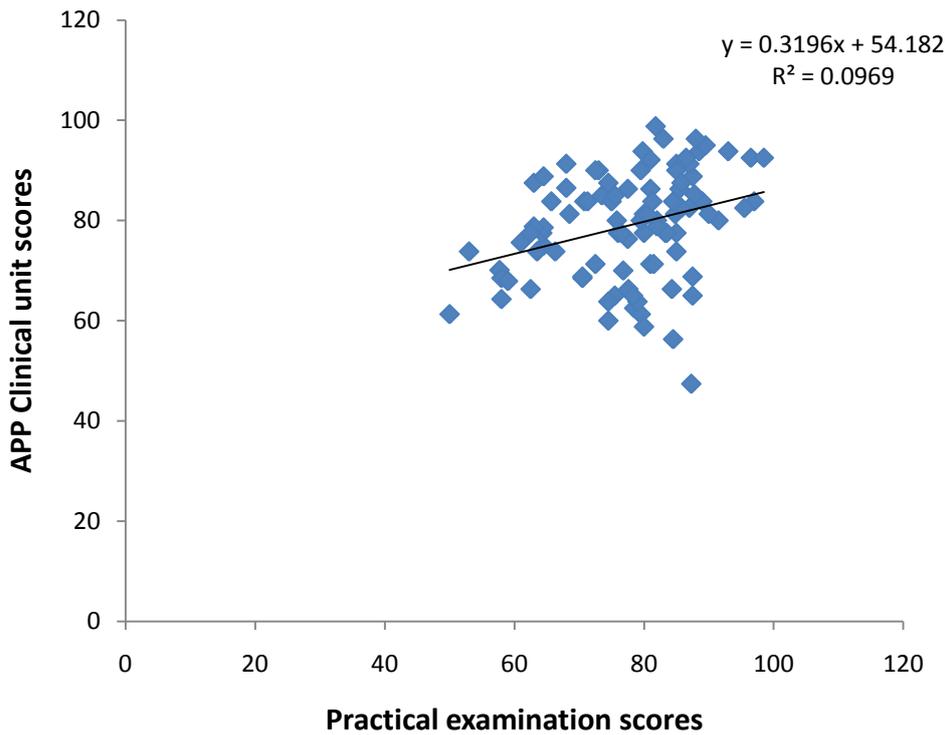


Figure 9.1: Scatter plot of total scores for orthopaedic practical exam and clinical unit

9.4.4 Characteristics of APP scores across six clinical units (n=57)

In the separate subgroup of 57 students followed longitudinally across all their placements one-way repeated measures ANOVA was conducted to compare APP scores at six consecutive time periods, Block 1: 0-5 weeks, Block 2: 5-10 weeks, Block 3: 10-15 weeks, Block 4: 15-19 weeks, Block 5: 19-23 weeks, Block 6: 23-27 weeks of clinical experience. The means and standard deviations are presented in Table 9.4.

Table 9.4: Descriptive statistics of six clinical blocks (n=57)

Weeks of clinical experience	Mean	SD
0-5 weeks	44.67	10.18
5-10 weeks	47.75	12.13
10-15 weeks	49.26	12.12
15-19 weeks	58.80	12.17
19-23 weeks	61.44	9.98
23-27 weeks	60.14	10.58

There was a significant effect for time [Wilks' Lambda= .28, $F(5,52) = 25.75$, $p < .0005$, multivariate partial eta squared = .71]. Thus student total APP scores increased with increasing hours of clinical experience within the sub sample (n=57) of students, as illustrated in Figure 9.2.

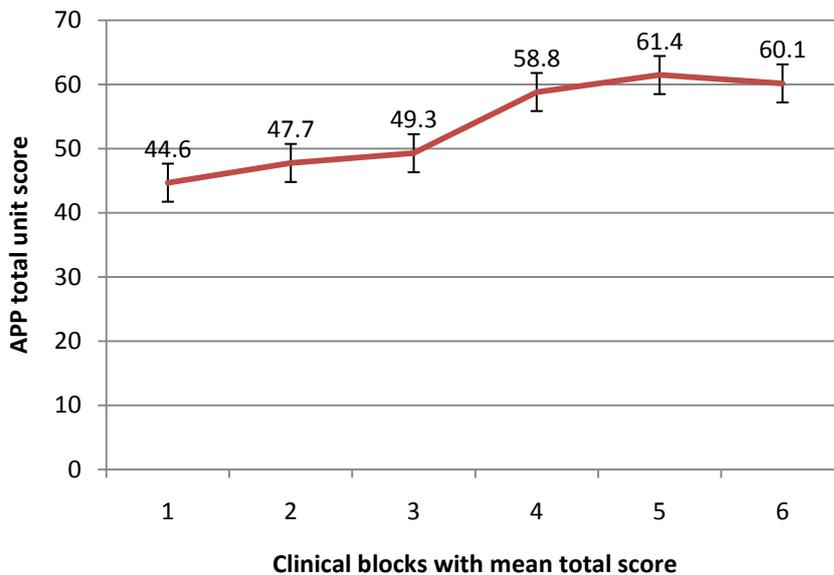


Figure 9.2: Change in mean APP scores across six clinical blocks

9.4.5 Factor analysis

The 20 items of the APP were subjected to principal components analysis (PCA) using SPSS version 14. (SPSS Inc.). Prior to performing PCA the suitability of data for factor analysis was assessed. Inspection of the correlation matrix revealed the presence of many coefficients of 0.3 and above. The Kaiser-Meyer-Olkin value was 0.977, exceeding the recommended value of 0.6 and the Bartlett's Test of Sphericity reached statistical significance ($p = .000$), supporting the factorability of the correlation matrix. Principle Component Analysis (PCA) demonstrated the presence of 1 dominant factor with an eigenvalue exceeding 1, explaining 61.2% of the variance as shown in Table 9.5.

Table 9.5: Component Matrix Field Test Two

Component	Total Variance Explained			Extraction Sums of Squared Loadings		
	Total	Initial Eigenvalues % of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	12.242	61.211	61.211	12.242	61.211	61.211
2	.931	4.947	66.158			
3	.719	3.596	69.753			
4	.648	3.240	72.994			
5	.527	2.634	75.628			
6	.490	2.452	78.080			
7	.455	2.276	80.356			
8	.424	2.120	82.477			
9	.404	2.021	84.498			
10	.351	1.757	86.255			
11	.337	1.685	87.940			
12	.329	1.646	89.586			
13	.307	1.535	91.121			
14	.294	1.471	92.592			
15	.277	1.383	93.975			
16	.263	1.316	95.291			
17	.253	1.264	96.555			
18	.247	1.233	97.788			
19	.228	1.141	98.928			
20	.214	1.072	100.000			

Extraction Method: Principal Component Analysis.

An inspection of the scree plot revealed a clear break after the first component (Figure 9.3). Using the scree test, it was decided to retain only one component for further investigation.

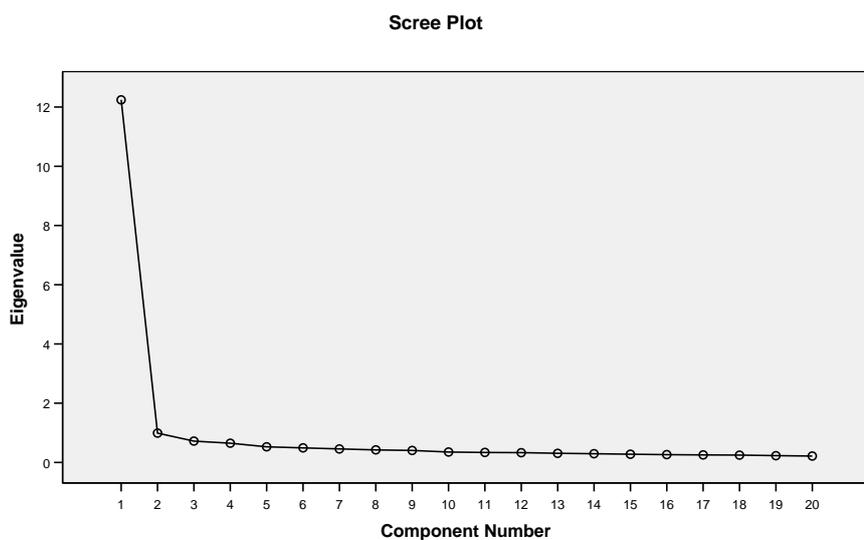


Figure 9.3: Scree plot Field Test Two.

Retention of one factor was further supported by the results of the parallel analysis (Pallant 2005) that showed only one component with an eigenvalue exceeding the corresponding criterion values for a randomly generated data matrix of the same size (20 variables x 644 respondents). As reported in Chapter Seven section 7.3.3 the Monte Carlo PCA for parallel analysis program by Watkins (2000) was used. This is demonstrated in Table 9.6. The results from parallel analysis support the decision from the screeplot to retain one factor.

Table 9.6: Factor analysis parallel analysis Field Test Two

Component no.	Actual eigenvalue From PCA	Criterion value from Parallel analysis	Decision
1	12.24	1.32	Accept
2	.93	1.26	reject
3	.71	1.22	reject
4	.64	1.18	reject
5	.52	1.15	reject

The component matrix (Table 9.7) shows the loadings of each item on the first component. SPSS uses the Kaiser criterion (retain all eigenvalues above 1) as the default. Table 9.7 shows all of the items load quite strongly (above 0.4) on the first component. This further supports the decision to retain only one factor. As only one component was extracted the solution cannot be rotated.

Table 9.7: Component matrix

Item	Component
	1
APP rating	.854
APP rating	.834
APP rating	.832
APP rating	.825
APP rating	.818
APP rating	.815
APP rating	.813
APP rating	.804
APP rating	.797
APP rating	.794
APP rating	.789
APP rating	.784
APP rating	.781
APP rating	.762
APP rating	.759
APP rating	.736
APP rating	.715
APP rating	.714
APP rating	.703
APP rating	.694

9.4.6 Rasch analysis: Model

Similar to Field Test One the Likelihood Ratio test was significant ($p < 0.001$) so the partial credit model was used. Similar to the first field test, in Field Test Two a large sample of data was available. As in the first field test data ($n=644$) were divided into two random samples, one ($n=326$) for model development and the other for model validation ($n=318$). The data were stratified and then randomised to optimise representation of completed APP instruments according to clinical area of the placement, level of student experience, facility type (hospital, non-government agency, community health centre, private practice) and university program (undergraduate, graduate entry).

9.4.7 Rasch analysis: Overall Model Fit

Sample 1 ($n=326$).

The Chi-Square Item-Trait Interaction statistic was 65.12 ($df= 80, p= 0.88$) with the Bonferroni adjusted alpha (α) value = .0025 (.05/20). The chi-square probability value of $p = 0.88$ indicated fit between the data and the model.

Validation sample 2 (n=318)

The Chi-Square Item-Trait Interaction statistic was 100.84 ($df= 80, p= 0.57$) with the Bonferroni adjusted alpha (α) value = .0025 (.05/20). The chi-square probability value of $p = 0.57$ indicated fit between the data and the model.

9.4.8 Overall Item and Person Fit

Sample 1 (n=326)

The residual mean value for items was -0.33 (SD 1.71), indicating presence of some misfitting items to the model. The residual mean value for persons was -0.26 (SD 1.19) indicating no misfit among the respondents in the sample.

Validation sample 2 (n=318)

The residual mean value for items was -0.32 (SD 1.73), again indicating some misfit of items to the model. Similarly the residual mean value for persons was -0.19 (SD 1.13) indicating no misfit among the respondents in the sample.

9.4.9 Individual Item and Person Fit

Sample 1 (n=326)

Similar to Field Test One, item 6 exhibited a positive item fit residual above 2.5 suggesting poor discrimination. None of the items exhibited a significant chi-square value (Table 9.8). Items 11 (Identifies and prioritises patient/client's problems) and 8 (selects and measures relevant health indicators and outcomes) displayed high negative fit residuals (-4.02 and -2.54) respectively (Table 9.8).

There were two people with positive fit residuals above 2.5. Investigation of these individual results revealed four instances of unexpected scoring on item 19 (evidence based practice), and one on item 6 (Communicates effectively – written communication). Deletion of these persons from analysis made no difference to overall model fit

Validation sample 2 (n=318)

Item 6 again exhibited a positive item fit residual above 2.5. None of the items exhibited a significant chi-square value (Table 9.8). Items 11 (Identifies and prioritises patient/client's problems) and 18 (undertakes discharge planning) displayed high negative fit residuals (-3.28 and -2.61) respectively (Table 9.8).

There was one person with a positive fit residual above 2.5. Investigation of the individual's results revealed one instance of unexpected scoring on item 3 (demonstrates ethical, legal and culturally sensitive practice), and one instance of unexpected scoring on item 19. Again, deletion of these cases made no difference to overall model fit.

Table 9.8: Individual item fit of 20 APP items to the Rasch model: Sample 1 (N=326) and sample 2 (n=318)
(Item order is from least to most difficult of the 20 items)

Sample 1 (n=326)								Sample 2 (n=318)							
APP item	Location	SE	FitResid	DF	ChiSq	DF	Prob	APP item	Location	SE	FitResid	DF	ChiSq	DF	Prob
1	-2.088	0.136	0.796	306.73	1.94	4	0.746	1	-1.824	0.128	1.104	280.98	5.765	4	0.217
3	-1.296	0.121	2.267	306.73	3.723	4	0.444	3	-1.516	0.119	1.726	280.98	4.587	4	0.332
2	-0.997	0.137	1.418	306.73	6.152	4	0.188	2	-0.532	0.124	-0.887	280.05	1.597	4	0.809
6	-0.647	0.121	4.479	306.73	13.939	4	0.007	5	-0.486	0.129	1.219	280.98	1.105	4	0.893
7	-0.455	0.116	-1.078	306.73	1.161	4	0.884	6	-0.466	0.112	3.671	280.05	0.665	4	0.955
4	-0.174	0.121	-0.358	306.73	3.856	4	0.425	7	-0.451	0.117	0.478	280.98	2.165	4	0.705
5	-0.154	0.114	0.46	306.73	1.759	4	0.779	4	-0.133	0.11	-2.121	280.98	1.462	4	0.833
20	-0.073	0.119	-1.85	306.73	3.346	4	0.501	20	-0.106	0.111	-1.863	280.98	5.841	4	0.211
14	-0.025	0.122	-0.539	305.79	1.537	4	0.820	14	-0.094	0.123	-0.724	280.98	8.107	4	0.087
15	0.286	0.114	-0.235	306.73	3.295	4	0.509	15	-0.011	0.119	1.108	280.98	2.286	4	0.683
16	0.297	0.115	-1.105	306.73	1.052	4	0.901	9	0.01	0.111	-0.14	280.05	3.503	4	0.477
18	0.401	0.122	-1.308	306.73	4.864	4	0.301	16	0.062	0.112	1.266	278.17	5.059	4	0.281
8	0.44	0.112	-2.54	306.73	6.308	4	0.177	18	0.11	0.119	-2.612	280.98	1.094	4	0.895
9	0.496	0.114	-2.166	306.73	3.993	4	0.406	8	0.158	0.111	0.741	273.49	8.757	4	0.067
11	0.508	0.114	-4.023	305.79	6.733	4	0.150	13	0.32	0.11	-1.285	278.17	3.389	4	0.494
13	0.509	0.113	2.14	304.85	3.857	4	0.425	19	0.321	0.112	-2.317	280.98	11.03	4	0.026
19	0.514	0.113	-0.178	304.85	2.162	4	0.706	11	0.719	0.111	-3.286	279.11	6.669	4	0.154
12	0.716	0.116	0.165	306.73	1.365	4	0.850	17	0.784	0.112	-1.008	280.98	8.001	4	0.091
17	0.845	0.115	-1.455	305.79	2.27	4	0.686	10	0.847	0.111	-0.61	280.05	7.732	4	0.101
10	0.896	0.115	-2.096	306.73	7.796	4	0.099	12	1.016	0.115	-0.827	280.05	3.024	4	0.553

Note: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies evidence based practice 20 = assesses risk

9.4.10 Threshold ordering of polytomous items

There were no disordered thresholds for any of the 20 items in either sample one or two. The threshold map for sample one is illustrated in Figure 9.4. An additional example of the ordering of thresholds is illustrated in Figure 9.5 in the category probability curves for item 4 (demonstrates teamwork) in sample two.

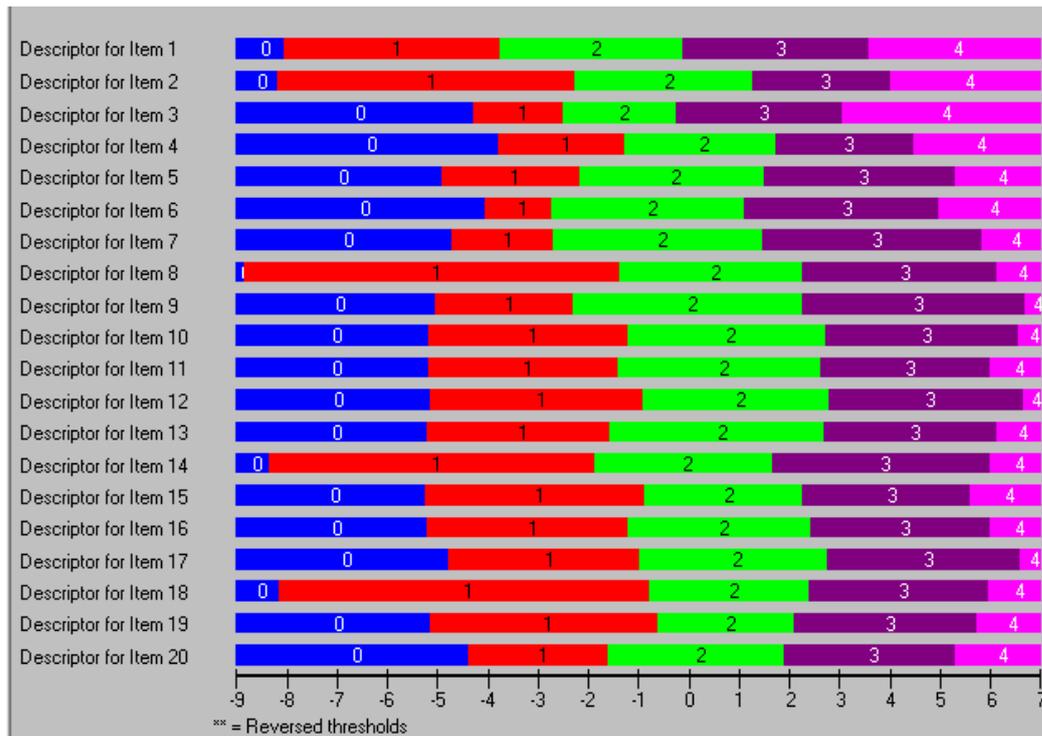


Figure 9.4: Threshold map of APP 20 items in sample 1 (n=326)

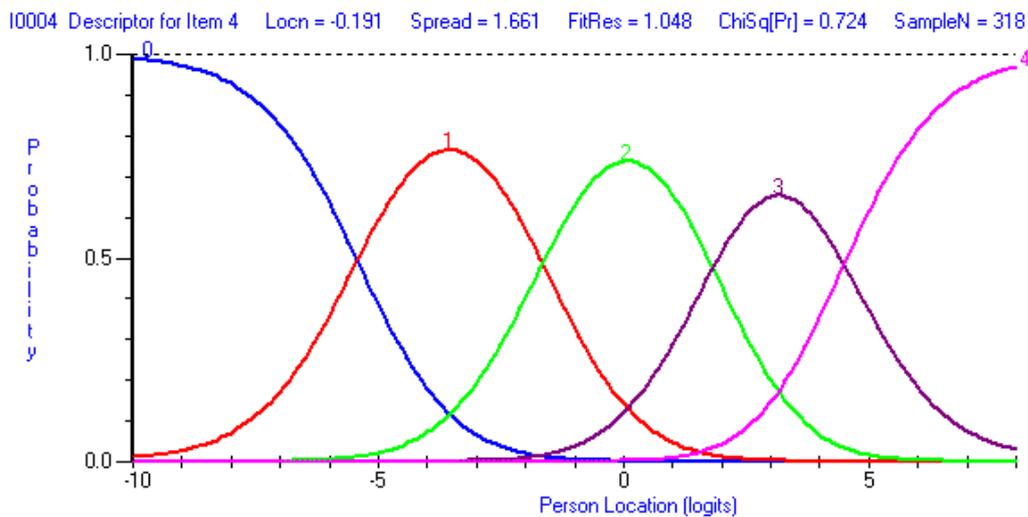


Figure 9.5: Category probability curves for item 4 in sample 2 (n=318)

Item 6 (demonstrates clear and accurate documentation) demonstrated a fit residual SD above +2.5 in both samples suggesting some misfit of this item which may be contributing to the overall misfit of items to the model. To investigate if the misfit of item 6 was contributing to the item misfit to the model, item 6 was removed from each sample and Rasch analysis repeated. When item 6 was removed and overall item fit re-examined, the residual mean value for items changed from -0.33 (SD1.71) to -0.33(SD 1.53) (Sample 1) and from -0.33 (SD1.73) to -0.32(SD 1.51) (Validation sample). In both samples, this indicated a modest improvement in the overall fit of items to the model as evidenced by the reduction in standard deviation.

9.4.11 Targeting

Visual inspection of the person-item threshold graphs, Figures 9.6 and 9.7, show the distributions of the students (top half of the graph) and item thresholds (bottom half of the graph) for the APP total score on a logit scale for both samples. Inspection of these person-item threshold graphs show that a majority of item thresholds correspond to the main cluster of persons (students).

Similar to Field Test One, there appears to be an even spread of items across the full range of student scores, suggesting effective targeting of APP items. At the far right hand end there are a few person abilities that have no equivalent item threshold difficulties that could differentiate their performance. These represent high performing students. The number of students who are performing at a level too low to be captured by the scale is negligible.

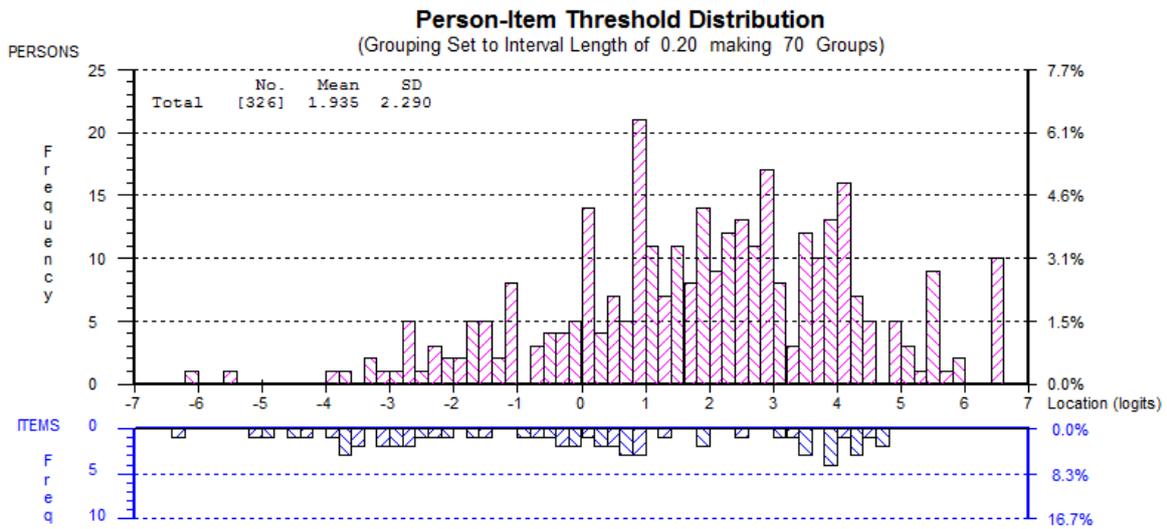


Figure 9.6: Person-item threshold distribution graph for sample 1 (n=326)

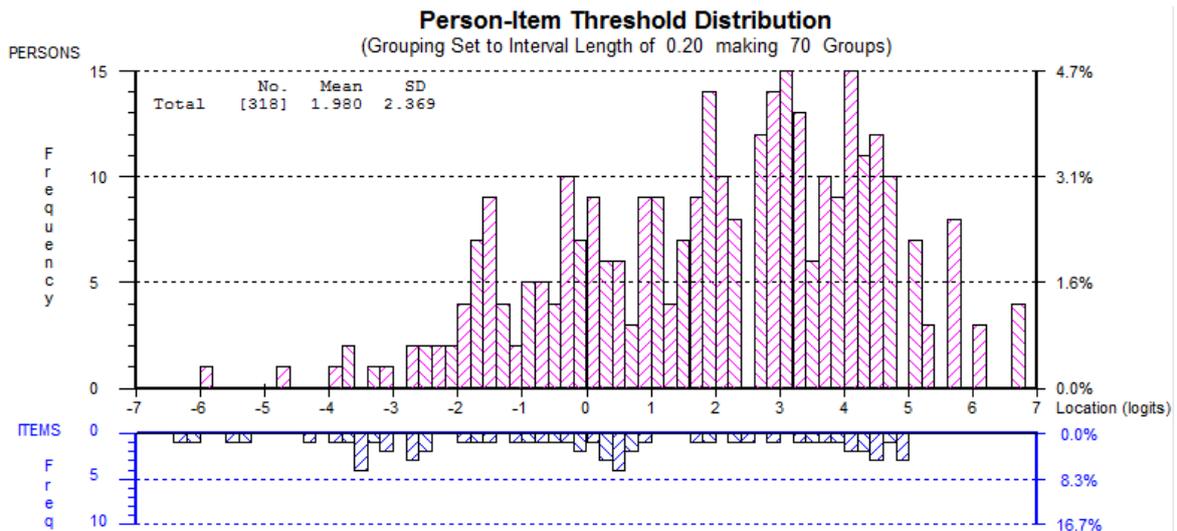


Figure 9.7: Person-item threshold distribution graph for sample 2 (n=318)

9.4.12 Hierarchy of item difficulty

The sequence or hierarchy of average difficulty of the 20 items on the APP for both samples are presented in Table 9.8, and graphically in Figure 9.8. In both samples, the first six items representing professional behaviour and communication were amongst the least difficult items whereas the most difficult items related to analysis and planning (items 12 and 10) and item 17 (progresses intervention appropriately).

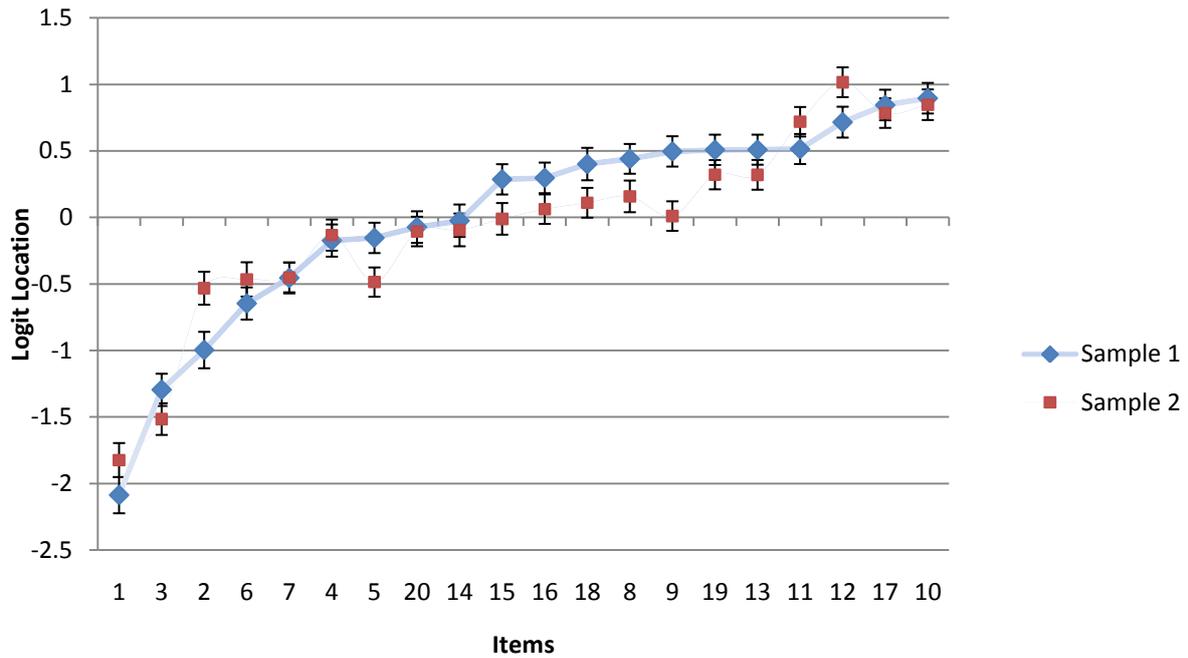


Figure 9.8: Logit location of APP items in two samples for Field Test Two

Note: Item 1 = understands client rights 2 = committed to learning 3 = ethical practice 4 = teamwork 5 = communication skills 6 = documentation 7 = interview skill 8 = measures outcomes 9 = assessment skills 10 = interprets assessment 11= prioritises problems 12 = sets goals 13= intervention choice 14 = intervention delivery 15 = effective educator 16 = monitors intervention effects 17 = progresses intervention 18= discharge planning 19 = applies EBP 20 = assesses risk

Figures 9.9 and 9.10 show the relationship between raw ordinal APP scores and person location logit scores for sample 1 and 2.

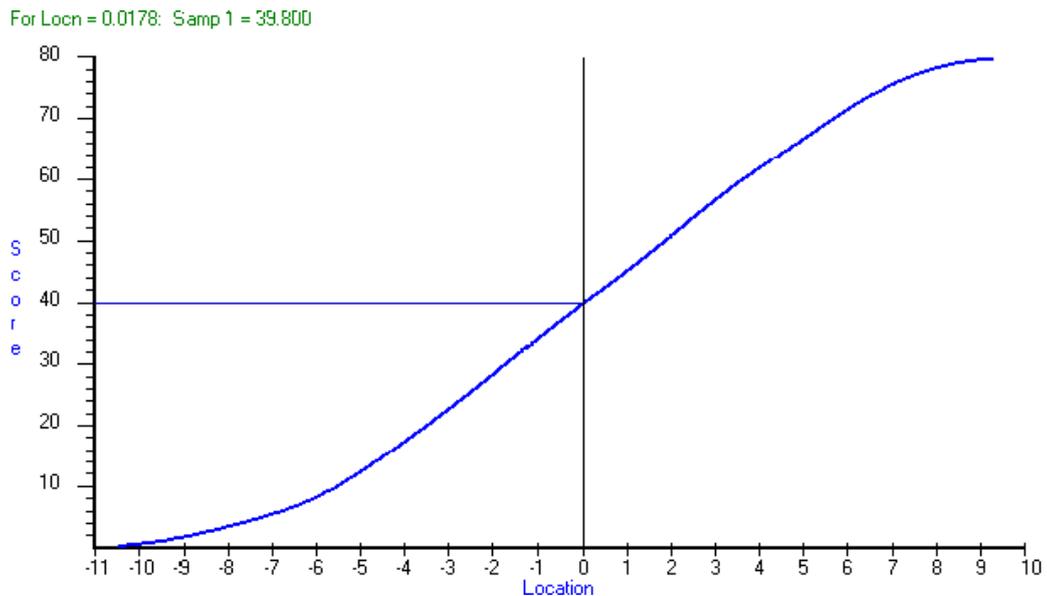


Figure 9.9: Plot of person logit location and raw APP score (sample 1)

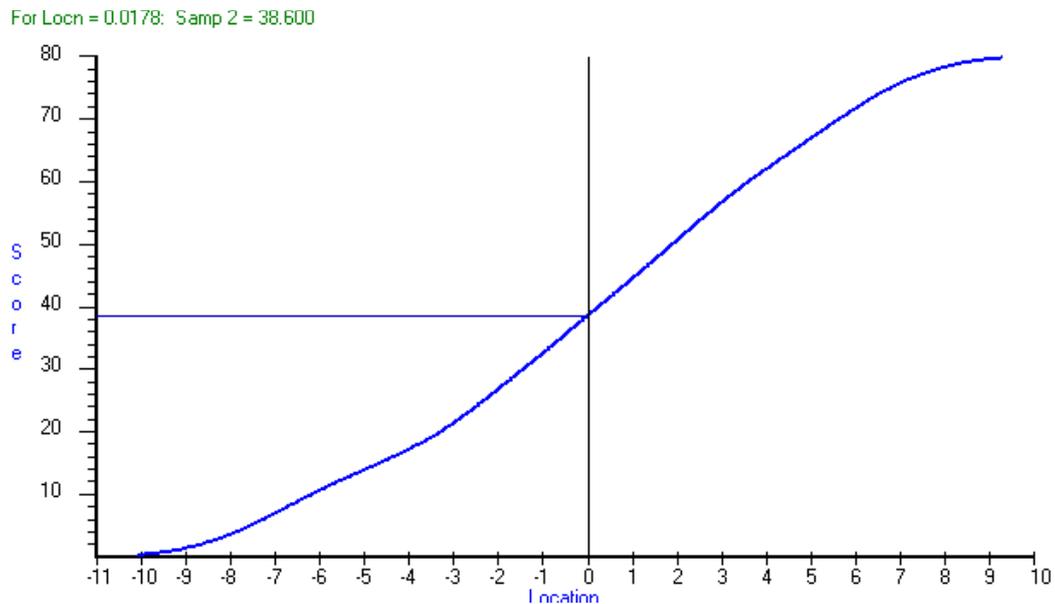


Figure 9.10: Plot of person logit location and raw APP score (sample 2)

9.4.13 Person separation index

Sample 1 (n=326) and validation sample 2 (n=318)

In sample one and two the PSI was 0.95 and 0.96 respectively indicating the ability to discriminate between 4 or more levels of performance.

9.4.14 Differential item functioning (DIF)

Similar to Field Test One, the presence of item bias was explored by analysis of DIF with a Bonferroni-adjusted p value of .0025 (.05/20). No significant DIF was demonstrated in either of the two samples for the following variables: student age, gender and level of clinical experience, clinical educator age, gender and experience as an educator, facility type, and clinical area. This indicates the APP item ratings were not systematically affected by any of these nine variables.

Student gender:

Unlike Field Test One, using the adjusted p value of .0025 in field test two, there was no DIF for item 6 for student gender in either sample. The results for sample 1 are shown in Table 9.9.

Table 9.9: Uniform and non-uniform DIF statistics for all APP items for student gender

Item	Uniform DIF				Non-uniform DIF			
	MS	F	DF	Prob	MS	F	DF	Prob
1	0.20819	0.18032	1	0.671421	0.73953	0.64052	4	0.634002
2	2.42391	1.99958	1	0.158423	0.79539	0.65615	4	0.622969
3	0.01891	0.01569	1	0.900394	2.15167	1.785	4	0.131845
4	0.16731	0.15414	1	0.694892	1.31372	1.2103	4	0.306533
5	1.78828	1.77416	1	0.18392	0.3854	0.38236	4	0.821194
6	3.66478	2.63748	1	0.105463	0.60939	0.43857	4	0.780702
7	0.08598	0.0999	1	0.752189	1.29849	1.50866	4	0.199679
8	0.86027	0.97678	1	0.323817	0.5802	0.65878	4	0.621124
9	0.00016	0.00022	1	0.988071	0.72016	0.99415	4	0.411021
10	0.34233	0.461	1	0.497699	1.03033	1.38751	4	0.238294
11	0.86133	1.00086	1	0.317942	0.52328	0.60806	4	0.657138
12	0.06971	0.07851	1	0.779534	0.43137	0.48586	4	0.746127
13	0.0006	0.0009	1	0.976186	0.9597	1.42562	4	0.225497
14	0.11872	0.156	1	0.693163	2.64323	3.4733	4	0.009623
15	3.15642	4.15804	1	0.042351	0.69844	0.92007	4	0.452553
16	2.11809	2.29837	1	0.130606	1.7112	1.85685	4	0.118104
17	0.0806	0.09596	1	0.756965	0.79854	0.95073	4	0.435023
18	1.58559	1.633	1	0.202347	3.56048	3.66693	4	0.011266
19	0.00468	0.00437	1	0.947273	0.93098	0.86956	4	0.482558
20	2.7437	3.06015	1	0.0813	0.36155	0.40324	4	0.806264

Student experience:

Similar to Field Test One, the level of student experience on clinical placements was examined for DIF by checking for the presence of DIF based on the number of weeks of clinical placement the student had attended prior to Field Test One. The level of student experience was coded as beginning (0-9 weeks prior experience), middle (10-19 weeks prior experience) and end (20-35 weeks prior experience). The individual item scores and the global rating scale (GRS) for each completed APP were examined for DIF based on time. None of the items or the GRS showed probability values exceeding the adjusted alpha value (.0025) in either sample.

8.3.15 Dimensionality

Analysis of the pattern of item loadings on the first extracted factor of the residuals shows that the residuals loaded in opposite directions on two subsets defined by positive and negative loadings on the first factor. Only those items with loadings greater or less than 0.3 were considered. Similar to Field Test One, some local dependence was evident, with four items showing positive residual correlations greater than 0.3 in both samples. The items

showing positive residual correlations were items 1 (demonstrates an understanding of patient rights and consent), 3 (demonstrates ethical, legal and culturally sensitive practice), 2 (Demonstrates a commitment to learning) and 5 (verbal communication).

The next step was to investigate if the person estimates (location values) based on scores that underpin each of these sets of items were significantly different using independent t-tests. A confidence interval for a binomial test of proportions was calculated for the observed number of significant tests. In sample one data, 24 cases out of 326 (7.3% or 0.073) had statistically different scores on each of the subsets of items. A confidence interval for a binomial test of proportions was calculated for this observed number of significant tests. The 95% confidence intervals around this estimate are calculated as expected: $16.3 (=0.05 \times 326)$, $14.7678 < 24.00012 < 33.252$ (Normal-z approx) or as a proportion of 0.073, $0.044 < 0.073 \text{ (obs)} < 0.10$. As the expected ranges contains the observed value 16.3 or .05, unidimensionality of the scale is supported (Smith, 2002).

This analysis was repeated for sample two data. In sample two data, 22 cases out of 318 (6.91% or 0.069) or expected: $15.9 (=0.05 \times 318)$, was $13.1334 < 21.99924 < 30.8778$ (Normal-z approx) or as a proportion of 0.069, $0.041 < 0.069 \text{ (obs)} < 0.096$. As the expected range contains the observed value 15.9 or .05, unidimensionality of the scale is supported (Smith, 2002).

8.3.16 Relationship of global ratings to person measures

A scatter plot of the student's global rating scale score and the students' overall Rasch location score was created for both samples. This indicated that as the students' overall global rating category increased as did their overall Rasch score (Figure 9.11).

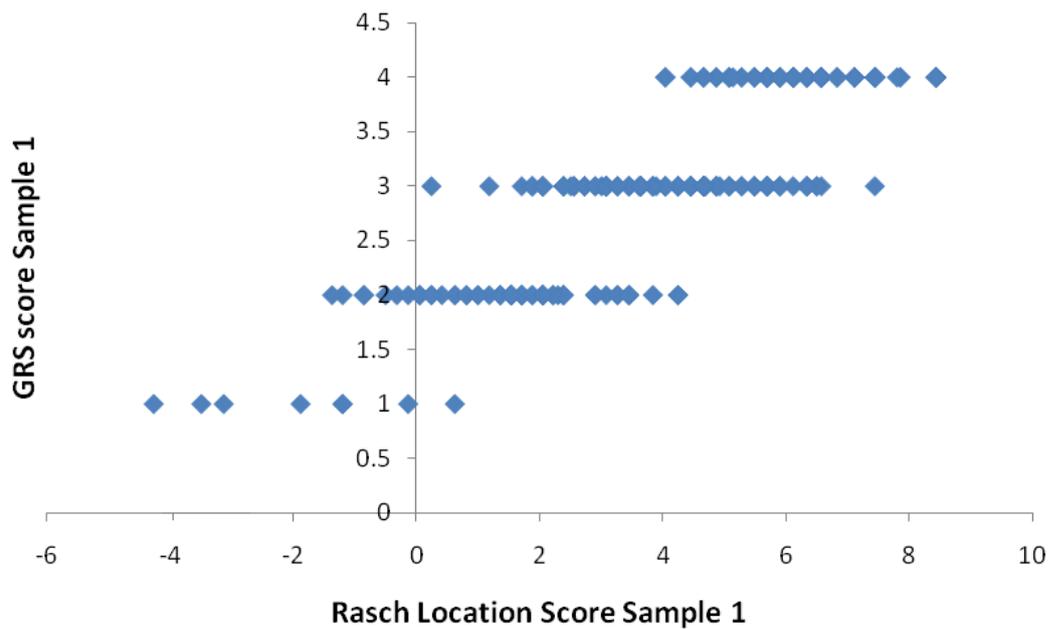


Figure 9.11. Scatter plot of Global rating scores (overall rating of competence) against Rasch person location logit score for sample 1 (n=326) in Field Test Two

9.4 Discussion

The results from Field Test Two were similar to those of Field Test One, with data from the second field test of the APP instrument consistent with the expectations of the unrestricted (partial credit) derivation of the Rasch model of measurement. These results confirm the validity of the 20 item APP as an instrument for measuring professional competence of physiotherapy students in the clinical environment. The discussion will focus on the similarities and differences between the results of both field tests.

Examination of the raw data revealed low levels of missing data with item 19 (application of evidence based practice in patient care) again being the most often not scored or scored as not assessed. However the rate had dropped from 4.4% in Field Test One to 1.2% in Field Test Two. The overall missing data and not assessed rate was 0.41%, half that of the rate found in Field Test One (0.85%), perhaps indicating that as a greater number of educators received training and became more familiar with instrument usage, they were more likely to score all items.

Examination of the hierarchy of average item difficulty shows a shift of item 19 from being the most difficult item (Table 7.5), to an average position in the top third of difficult items.

Additionally, item 19 no longer demonstrated item misfit in Field Test Two. The hierarchies of both samples in Field Test Two revealed that analysis and planning (critical thinking), goal setting and selection and progression of interventions were the most difficult items. Apart from the position of item 19 as the most difficult item in the first field test, the sequencing of items in both field tests is closely aligned with educators' experience of domains of practice they observe to be more difficult for students to master.

While the data demonstrated overall fit to the Rasch model for both participant samples, item fit residual values indicated the presence of misfitting items. In Field Test Two as in the first field test, item 6 (written communication) continued to show misfit to the Rasch model. Investigation of DIF did not demonstrate the presence of DIF for student gender in either sample in field test two. Further investigation of item misfit by removing item 6, achieved only a modest improvement in item fit residual values as evidenced by the reduction in standard deviation. Despite some improvement in overall item fit to the model, removal of item 6 is not justified given that written communication is part of the current APC standards and represents an essential aspect of professional competence as a physiotherapist. As discussed in Chapters 6 and 8, focus group and clinical educator training workshop data revealed division between educators on the acceptable minimum standard for written communication. The discrepancy in minimum standard occurred primarily in relation to writing up patient/client clinical notes rather than discharge summaries and letter writing to referring doctors or referral to other health professionals. Some educators considered the taking of notes during the patient interview to be essential, while others felt note taking impacted on effective communication with the patient and required students to write up notes on completion of the interview or during a scheduled break. Where there were facility based clinical assessment proformas, students were generally expected to complete these during the patient interview. The clinical area also influenced educator's expectations. In the acute inpatient hospital ward setting, the expectation was to write up multiple chart entries after assessing and treating several patients. The availability of patient charts in a busy ward environment was cited as the main reason underpinning this approach. In the outpatient setting, patient notes were often completed during the patient interview. Overall the expectations of educators varied and were often not discussed with the student prior to commencement of the clinical placement. This signals an opportunity for practice

standardisation and highlights the advantages of using a standardised instrument that uncovers issues such as this one and enables reflection by the profession. Overall qualitative data suggests that variable expectations of minimum acceptable standards of written communication are an issue for clinical educators and the broader physiotherapy profession. Further research on this item and how it is being interpreted and scored by educators is warranted.

In both field tests no significant DIF was demonstrated for the variables student age and experience, clinical educator age, gender and experience as an educator, university, or field of practice. This indicates the APP item ratings were not systematically affected by any of these variables and supports nationwide use of this instrument across all clinical areas, facilities and universities.

In both field tests, targeting of the APP showed adequate coverage of thresholds across the whole construct of professional competence, and the scale exhibited high reliability (PSI = .95) with no disordered thresholds, indicating a highly discriminative scale. These data allow educators to be confident in the ability of the APP scores to differentiate between students with varying levels of ability. In the pilot trial and both field tests some local dependence was evident with items 1 – 5 showing positive residual correlations in both trials. There is a logical relationship between these items as they all relate to the professional behaviour and communication domains of practice. Similar to the first field test, Field Test Two demonstrated that the APP was robust when tested against the assumptions of the Rasch measurement model, with the independent t-test analysis supporting the assumption of unidimensionality. Factor analysis of Field Test Two raw data determined the presence of one dominant factor explaining 61% of the variance also supporting the concept of unidimensionality of the APP.

High negative item fit residuals evident in several items in both field tests, suggests some redundancy or overdiscrimination in some items in the samples. A degree of redundancy in these items is considered acceptable as they are essential components of professional competence requiring assessment and there is insufficient evidence to support their removal on statistical grounds (Streiner & Norman, 2003).

While the results of the Pearson product-moment correlation for the four different assessment tasks examining orthopaedic physiotherapy knowledge and practical skills lend support to the validity of the APP instrument based on relation to other variables, it is far from definitive. Calculation of the variance (r^2) shows that performance on practical skills predicted only 10% of APP total score. As Hobart et al (2007) state the limitation of this approach to establishing validity is that, to show that scores from an instrument do not correlate highly with measures of a dissimilar construct or correlate highly with measures of a similar construct, tells us nothing about what the scale actually measures. This approach informs us only that the two are related or not. However, when this source of evidence is combined with evidence of face validity, alignment to required practice standards and widespread stakeholder support that the items are comprehensive, an argument is built that it is likely that the scores measure the construct that the profession calls 'entry-level professional competence'. Validity evidence can be assembled from a wide variety of sources, and ongoing evidence should be sought and assembled as APP scores are collected in concert with other related and unrelated assessment outcomes (Chapter Eleven).

A one-way repeated measures ANOVA demonstrated that student total APP scores increased significantly with increasing hours of clinical experience within a sub sample (n=57) of students. While this supports the hypothesis that increasing time spent in the clinical environment is accompanied by higher levels of professional competence, hours alone may not be the sole predictor of higher levels of competence. Other potential factors instrumental in the development of professional competence may include the effect of curriculum, the specific nature of the student experience in different workplace environments and the skill of the educator in facilitating student learning (McAllister, et al., 2010). Inspection of Figure 9.2 also reveals that student scores appear to plateau in blocks five and six. As these are the final two clinical blocks it could be that these students are working at or close to the level at which they will graduate and until they experience unsupervised practice in the work-place and its associated responsibilities, they may change little with additional educator supervision.

Overall from both field tests there is evidence that time seems to be accompanied by increasing APP scores, but we are not sure to what extent clinician expectations of

performance drive this relationship. A future study where students are rated by clinicians blind to the level of previous experience of the students would help to differentiate educator bias from observable evidence of competence. The relationship between student work-based learning and development of professional competence would benefit from further research.

Clinical educators were asked to give students a global or overall rating of competence on a global rating scale on the APP. This rating was included with a view to evaluating the utility of a global rating as an indicator of the level of competence students had attained and to examine the relationship of this global assessment of professional competence with item ratings. Similar to Field Test One results, the GRS mean scores increased as the level of clinical experience increased.

Visual inspection of the scatterplot of GRS and Rasch scores also showed that, as the students' overall rating category increased, so did their overall Rasch score. It is evident that there is overlap between the various categories of global ratings and associated person location measures. This suggests that there is not a discrete one-to-one correspondence between the degree of professional competency defined by summarising all ratings on the items and converting them into a Rasch score, and the overall global rating representing the clinical educator's judgement of professional competence. There are several possible explanations for this. The APP scores have some error associated with their measurements (this is described in Chapter Ten). The GRS is also imperfect, as it is based on clinicians' global impressions, and these vary with educator experience and expectations. It is also possible that certain student attributes affect global rating, e.g. students with winning personalities may achieve high GRS scores, despite measureable opportunities for improvement evident in specific item scores. Different items may contribute differentially to an impression of overall competence, e.g. it is possible that students who do not attend to strategies to minimise risks in the workplace might be rated as poor on a GRS, despite high specific item scores, as clinician educators have repeatedly described the importance placed on the risk management item. Despite the imperfect relationship between APP and GRS scores, this plot provides further validity evidence of anticipated changes over time concerning the target construct of professional competence (Wolfe & Smith, 2007b).

Similar to Field Test One data (Chapter Seven) conversions from raw scores to Rasch scores can be provided, but this adds a layer of complexity to calculating the student's final score that it appears can be avoided due to the almost perfect linear relationship shown in Figures 9.9 and 9.10. These Figures demonstrate a linear relationship, with slight flattening at scale extremes which indicates that raw scores can be used with confidence as though they were interval, unless scores are at the extremes.

9.5 Actions arising following Field Test Two

Evaluation of Field Test Two qualitative data (Chapter Eight) and quantitative data in this Chapter, lead to a number of changes being made to the APP (version 4) instrument. A summary of the modifications is presented in Table 9.10. The final amended APP instrument (version 5) is provided in Appendix 9.2. Following completion of the pilot trial and two field tests a considerable amount of qualitative and quantitative data had been assembled. To assist the reader to assimilate the data, a summary of the important findings is presented in Table 9.11.

9.6 Chapter Summary

Analysis of the quantitative results indicated that the APP data had adequate fit to the chosen measurement model (Rasch Partial Credit Model), the Person Separation Index demonstrated the scale was internally consistent discriminating between four groups of students with different levels of professional competence, the items were targeting the intended construct (professional competence) and the instrument demonstrated unidimensionality. Further research on how educators are interpreting and scoring the written communication item is required. Field Test Two provides data supporting the validity of the APP instrument scores and the findings from the pilot trial and first field test.

Table 9.10: Modifications to APP (version 4) following Field Test Two

Requested modifications to APP (v4)	APP (v 4) used in field test 2	APP (v 5) – final instrument
<p>Reword item 3. NZ requested that cultural sensitivity be made more obvious</p> <p>Performance indicator on cultural sensitivity to be added in</p>	<p>Item 3</p> <p>Demonstrates practice that is ethical and in accordance with relevant legal and regulatory requirements</p>	<p>Item 3</p> <p>Demonstrates ethical, legal and culturally sensitive practice</p> <p>Additional performance indicator</p> <ul style="list-style-type: none"> Practises sensitively in the cultural context
<p>Reword item 6 to differentiate written communication more clearly from verbal/non verbal communication in item 5</p>	<p>Item 6</p> <p>Communicates effectively and appropriately – Written</p>	<p>Item 6</p> <p>Demonstrates clear and accurate documentation</p>
<p>On the rating scale, replace n/a with full wording, not assessed to prevent confusion with not applicable.</p>	<p>0 1 2 3 4 n/a</p>	<p>0 1 2 3 4 not assessed</p>
<p>GRS: add in wording that GRS is not to be completed at mid unit.</p>	<p>In your opinion as a clinical educator, the overall performance of this student in the clinical unit was:</p> <p>Not adequate <input type="checkbox"/> Adequate <input type="checkbox"/></p> <p>Good <input type="checkbox"/> Excellent <input type="checkbox"/></p>	<p>DO NOT COMPLETE GLOBAL RATING SCALE AT MID UNIT</p> <p>In your opinion as a clinical educator, the overall performance of this student in the clinical unit was:</p> <p>Not adequate <input type="checkbox"/> Adequate <input type="checkbox"/></p> <p>Good <input type="checkbox"/> Excellent <input type="checkbox"/></p>

Legend: NZ: New Zealand, GRS: global rating scale; v: version.

Table 9.11: Summary qualitative and quantitative data following pilot and field testing

Themes / criteria	Qualitative data results – CE and S (Chapters 5, 6 and 8)	Quantitative data results (chapters, 4, 7 and 9)
Items	<ul style="list-style-type: none"> • Overall items comprehensively cover all aspects of clinical practice (CE & S) • Item 6 (written communication) – CEs appear to have inconsistent interpretation of passing standard. • Item 19 (EBP) – some CEs have difficulty assessing this item. • Item 18 (discharge planning) – initially misunderstood by educators in FT1. • Applicability of APP across placement settings, clinical areas, gender and experience level of CE, University, 	<ul style="list-style-type: none"> • Overall fit of the data to the Rasch model • The residual mean value for items indicated presence of some misfitting items to the model. • Items 6 & 19 : misfitting items in FT1 • Item 6: misfitting item in FT2 • Item 6: DIF for student gender. Male students consistently score lower than female students • No DIF for following variables: student age, clinical educator age, gender and experience as an educator, University, clinical area, facility type. This indicates the APP item ratings were not systematically affected by any of these variables <p><i>Student feedback questionnaire:</i></p> <ul style="list-style-type: none"> • Items easy to understand mean(SD)= 3.9(0.7)
Performance Indicators (PIs)	<ul style="list-style-type: none"> • Overall PIs comprehensively cover all aspects of clinical practice (CE & S) • PIs: Improve consistency when providing formative feedback (CE & S) • PIs: Assist students to self assess areas of performance requiring improvement 	<p><i>Student feedback questionnaire: 5 point scale (1 = Strongly Disagree and 5= Strongly Agree)</i></p> <ul style="list-style-type: none"> • PIs useful to assess own performance mean(SD)= 3.9(0.7) <p><i>CE feedback questionnaire: 5 point scale (1 = Strongly Disagree and 5= Strongly Agree)</i></p> <ul style="list-style-type: none"> • PIs useful = M (SD) 4.1(0.7) • PIs easy to understand= M (SD) 4.1(0.6)
Scoring systems – item rating scale and GRS	<ul style="list-style-type: none"> • Continued requests from some CEs and students for additional scoring category for items and on GRS. • Some CEs request splitting APP into 2 subscales – professional behaviour and physiotherapy specific skills • Request to weight more difficult items • CEs using full range of the scale from 0 – 4 • Infrequent use of not assessed 	<ul style="list-style-type: none"> • No disordered thresholds for 20 items and PSI=0.96 representing the ability to discriminate between 4 or more levels of performance. This suggests the scoring scale is functioning appropriately • Rasch analysis FT 1 & 2 show APP to be unidimensional. PCA FT 1 & 2: show APP to have one dominant factor suggesting no requirement to split instrument. <p><i>CE feedback questionnaire:</i></p> <ul style="list-style-type: none"> • Confident using 0 – 4 rating scale M (SD)=4.0(0.6) • Confident using GRS M (SD)=4.0(0.8) <p><i>Student feedback questionnaire:</i></p> <ul style="list-style-type: none"> • Rating on GRS was a fair indication of performance M (SD) 4.0±1.3 • Rating on 20 items were a fair indication of performance M (SD) 3.9± 1.2

Themes / criteria	Qualitative data results – CE and S (Chapters 5, 6 and 8)	Quantitative data results (chapters, 4, 7 and 9)
Passing standard	<ul style="list-style-type: none"> • Some CEs having difficulty with use of entry level as passing standard for early clinical units • GRS used consistently by CEs during summative assessment. Inconsistent use of GRS at mid unit: some educators scoring the GRS in relation to the student's prior experience rather than against entry level performance as instructed. • Video exemplars of students demonstrating a passing performance requested as a strategy to assist in training of both students and educators (DVDs now developed) 	<ul style="list-style-type: none"> • Inspection of person item threshold graphs show an even spread of items across the full range of student scores, suggesting effective targeting of the APP items.
Training Workshops Resource manual	<ul style="list-style-type: none"> • CES welcomed consistent approach to training in assessment offered during FT1 & 2 • CEs and students requested continuation of training to maintain standardised use of APP • Consensus on Resource manual as a comprehensive guide to effective assessment (CEs and S) • Need for training of CEs in areas of EBP, and written communication 	<p><i>CE feedback questionnaire:</i> Resource manual was comprehensive M (SD)= 4.3 (0.6)</p> <p><i>Student feedback questionnaire:</i> Information about the APP was adequate M (SD)= 4.1(0.7)</p>
Format of instrument	<ul style="list-style-type: none"> • Clear consensus from CEs and students on layout/format of APP • Preference by end of FT2 for APP to be available electronically as well as paper based format 	<ul style="list-style-type: none"> • Low levels of missing data for both FT1 & 2 (0.2%) <p><i>CE feedback questionnaire:</i></p> <ul style="list-style-type: none"> • Pls easy to understand M (SD)=4.1(0.6) • Scoring rules helpful M (SD)=4.1(0.7) • Pls useful M (SD)= 4.1(0.7)
Feedback	<ul style="list-style-type: none"> • Consensus that APP effective in providing specific, well targeted formative feedback (CEs and students) 	<p><i>Student feedback questionnaire:</i></p> <ul style="list-style-type: none"> • Pls useful to assess own performance M (SD)= 3.9(0.7) <p><i>CE feedback questionnaire:</i></p> <ul style="list-style-type: none"> • Pls useful when providing feedback= M (SD) 4.1(0.7)
Acceptability	<ul style="list-style-type: none"> • APP acceptable to both CEs and students for use in clinical environment Time taken to complete acceptable to CEs in both FT1 M (SD)=21.65 (13.3) mins; FT2 M (SD)= 29.04(19.3) mins 	<p><i>CE feedback questionnaire:</i> APP practical for use in the clinical environment= M (SD) 4.1(0.6)</p> <p><i>Student feedback questionnaire:</i> APP practical for use in the clinical environment = M (SD) 4.1(0.7)</p>
Others Standardisation On-line instrument	<ul style="list-style-type: none"> • One assessment instrument for all universities assists standardisation of assessment practices and training • CE: Shifted preference toward completion of APP on-line by end of FT2 	<ul style="list-style-type: none"> • The sequence or hierarchy of average difficulty of the 20 competencies on the APP fit closely with the experience of clinical educators regarding items they observe to be more difficult for students to master

Legend: CE: clinical educator; S: student; FT1: Field Test One; FT2: Field Test Two; PSI: person separation index; DIF: differential item functioning; GRS: global rating scale; Pls: performance indicators; EBP: evidence based practice; APP: Assessment of Physiotherapy Practice instrument; M: mean; SD: standard deviation; mins: minutes

10.Chapter Ten: Reliability

10.1. Introduction

Reliability is the extent to which assessment of performance yields relatively consistent results across occasions, contexts and assessors (Baartman, et al., 2007). Like validity, it is a property of the score and not the instrument itself (American Educational Research Association, 1999). Reliability evaluation provides a method for estimation of the amount of error, random and systematic, inherent in measurements. Reliability is dependent on the characteristics of the test, the conditions of administration, the group of examinees and the interaction between these factors (Streiner & Norman, 2003; Wolfe & Smith, 2007a).

Evidence from multiple sources can be accumulated to establish the likely validity of interpretations made based on the instrument scores. Measurement reliability constitutes one source of relevant evidence. An instrument that yields scores with inadequate consistency in different circumstances, when the underlying construct (in this case, professional competence) is unchanged, would be of limited value no matter how sound other arguments are for its validity. Hence the assessment of reliability of APP scores reported in this chapter informs the subsequent discussions regarding both reliability and validity for assessment of professional competence.

While repeated testing of the same student under the same conditions in the authentic practice environment is rarely feasible in performance based assessment, the consistency with which different assessors rate the performance of different students (inter-rater reliability) is very relevant. Since inter-rater reliability contains all the sources of error contributing to intra-rater reliability, plus differences that arise in decisions made by different observers, demonstration of adequate inter-rater reliability is sufficient evidence of adequate intra-rater reliability (which is typically more reliable) (Streiner & Norman, 2003; Wilson, 2005).

Assuming that there is a true value for professional competence, two sources of error in ratings are of interest. One is the random variation in scores when the same underlying

professional competence is assessed by independent assessors; the other is the systematic variation in scores. The latter may result, for example, from assessors with different expectations of entry level competence for individual items on the APP, or from different circumstances within which the student is assessed that enable or restrict a view of student competence. Systematic variation is of interest because it may be possible to trace the source of errors of this nature and correct them with methods such as standardised training of assessors, or adjustment of grades for areas of practice where higher level skills are typically expected, (e.g. critical care wards). Random errors are, by their nature, unpredictable. They need to be estimated and allowed for in score interpretation (Rankin & Stokes, 1998).

Different correlation indices can be computed, some of which (e.g. Pearson's r , ICC 3,1), ignore systematic error and provide data on the degree to which ranking of students is consistent (Rousson, Gasser, & Seifert, 2002). If these indices are entered into formulas for estimating metricated indices of reliability, the resultant error in estimation around an obtained score reflects only random error. Systematic error can be tested for using simple tests for differences in repeated measures such as a paired t -test. Other correlation indices (eg ICC 2,1), pool both systematic and random error. This might be sensible if no attempt was planned to identify and limit systematic error, or if there were credible reasons why any observed systematic error should be constant across the predicted applications of the test. If there is little systematic error and test-retest differences are not significant, metric error estimates derived from correlational indices will be similar regardless of the type of coefficient used.

10.1.1 Establishing Reliability

The approach used to assess reliability depends on the intended application of the assessment procedures. In the typical assessment of professional competence of physiotherapy students, raters are drawn randomly from a pool of educators. Hence this reliability study was designed so that 60 different educators formed 30 independent pairs of assessors. Each student was assessed once by a unique pair of educators. The grades

attributed to student performance therefore provided reasonable representation of the grades that would be awarded in assessments in the authentic practice environment.

There has been considerable debate regarding the best approach to estimating and describing reliability (Bland & Altman, 1986; McGraw & Wong, 1996; Rankin & Stokes, 1998; Rousson, et al., 2002; Shrout & Fleiss, 1979; Streiner & Norman, 2003).

The most commonly used methods include intraclass and Pearson's product-moment correlation coefficients and the Bland and Altman method (Bland & Altman, 1986; Downing, 2004).

Shrout and Fleiss (1979) described six forms of the Intraclass Correlation Coefficient (ICC), later expanded to ten forms by McGraw and Wong (1996). The various ICCs can be calculated from mean square deviations (MSD) derived from a within-subjects, single-factor (repeated measures) ANOVA. The ICC can theoretically vary between 0 and 1 where an ICC of one represents perfect agreement of repeated scores and 0 indicates no agreement. Weir (2005) proposed four issues requiring resolution in selecting an appropriate ICC.

Issue 1: One or Two –way model

The one-way ANOVA combines all sources of error in MSD estimates, while in two-way analysis MSDs are estimated separately for systematic and random error. In this study there was no reasonable expectation that error would be systematic, as the first and second raters (educators) were always different, and the order in which they reported results was a random event. Even if systematic error was observed, there is no plausible reason why it should be expected to occur if the study were repeated. Therefore it was decided to treat all the error as if it were random error, combining any apparent systematic difference in test one/test two mean scores with observed random error. A two way ANOVA model provided the MSDs required to compute all potential sources of error.

Issue 2: Fixed or random effects model

In a random-effects model, raters are considered to be a random sample from a larger population of potential raters. This would be the typical circumstance in the practice environment in which students were assessed. In a fixed effects model the raters who are

assessed are the only raters of interest (a relatively rare event). Hence in this study, analysis using a random effects model was planned.

Issue 3: Include or exclude systematic error

The most compelling reasons to partition systematic and random error are the potential to identify the source of the systematic error and reduce it (e.g. providing more practice to raters) and to enable reflection on the likely generalisability of estimates of systematic error to other raters or test conditions (e.g. systematic error attributable to time between test and retest might only be expected under comparable test schedules). As argued with respect to Issue 1, systematic error was not expected given the random selection of raters and the random order in rating. Hence any observed systematic error was treated as if it were a random event in the sampled data and pooled with random error in analysis.

Issue 4: Whether to use single or mean scores for repeated measurements

Whether one chooses to estimate the reliability of a single measure or the reliability of the average of a number of measures should be determined based on the likely application of the test in the authentic application environment. If the use of the average of repeated assessments is impractical in practice, then the reliability of a single measurement is the statistic of interest. In the authentic practice environment, assessors are restricted by time and workload and multiple summative assessments rarely occur and would not be favoured by educators. Hence a single measurement provided the data that was analysed for reliability.

Combining these four considerations, the Intraclass Correlation Coefficient 2,1 (two-way random effects model) appeared to be the appropriate ICC. There is, however, no single reliability coefficient that adequately conveys all relevant information about reliability. The standards provided by the American Educational Research Association (1999) state that instrument developers are obligated to provide sufficient data to enable those using the instrument to make informed judgments regarding whether scores are precise enough for the users' intended interpretations. When considering intraclass correlation coefficients, Streiner and Norman (2003) recommend that a coefficient of 0.75 and above is a minimal requirement for a useful instrument. Landis and Koch (1977) similarly recommended that

coefficients between 0.61 and 0.80 represented substantial strength of agreement between measurements, whereas Portney and Watkins (1993) recommended 0.90 for making decision about individual subject scores. Correlation coefficients provide information on the utility of measurements to differentiate between different individuals and are an index of consistency in ranking order. This would be useful for tests such as IQ tests, where the error in establishing a hierarchy of scores for a group may be of importance to the examiner. What is lacking in this approach is that it does not provide information about the magnitude of error (expressed in the scale units of measurement) associated with a single application of the test, or repeated applications of the test under conditions when it is reasonable to expect that the underlying construct has not changed (Keating & Matyas, 1998; Streiner & Norman, 2003; Weir, 2005). In addition, as ratios of within subject variance attributable to different raters to total variance (that includes variance in scores for individuals) correlation coefficients are strongly influenced by the range of scores obtained by the sample. Where range is attenuated (eg all students perform well), correlational indices can be low even when error is acceptably small. Similarly, where there are large differences between the performance of individuals, high correlations for test -retest data can occur despite unacceptably large error margins. Hence the (quite variable) recommendations regarding what magnitude of correlation constitutes evidence of adequate reliability needs to be considered a 'rule of thumb' and examined in the context of other indicators of measurement stability. Metricated estimates of error help illuminate the potential utility of scores for their intended application.

The standard error of the measurement (SEM) describes one standard deviation of the typical error associated with a single rating. Because it refers to error in a single test score, it is calculated using Pearson's correlation coefficient, which does not include systematic retest variance, and subsequently describes only the magnitude of random error typically associated with measurements. The lower the value of the SEM, the smaller the random error associated with the measurement (Rankin & Stokes, 1998). The SEM is one of several approaches to quantification of the precision of individual scores on an instrument or test. The formula for calculation of the SEM is presented in *Equation 1*.

$$SEM = SD\sqrt{1-r}$$

Equation 1

Where:

r = Pearson's correlation coefficient for test-retest scores

SD = the standard deviation of raw scores obtained in the repeated measures study.

The SEM can be used to calculate a confidence interval (CI) around an observed score as shown in this equation:

$$X_0 \pm z(SEM) \quad \text{Equation 2}$$

Where:

X_0 = the observed score

Z = the value from the normal curve associated with the desired CI (1.64 for 90% CI; 1.96 for 95% CI).

The Minimal Detectable Change (MDC) provides an estimate of the magnitude of change that must be seen in a measurement to exceed the anticipated measurement error and variability (de Vet, et al., 2006; Ries, Echternach, Nof, & Gagnon Blodgett, 2009; Stratford, 2004) and conclude that real change has occurred. The MDC_{90} provides the error estimate that represents expected score variation (in the absence of real change) in 90% of cases (Equation 3).

$$MDC_{90} = SD_{Diff} \times 1.65 \quad \text{Equation 3}$$

Where:

SD_{Diff} = the standard deviation of difference score for test -retest scores collected under conditions where no real change is considered likely

1.65 is the z-score that defines the limit that includes 90% of expected differences. This is replaced by the appropriate t value when estimations are based on small samples.

The MDC_{90} can alternatively be calculated using the SEM (Equation 4)

$$MDC_{90} = \sqrt{(SEM_1^2 + SEM_2^2)} \times 1.65$$

Equation 4

Where;

$SEM = SD\sqrt{1-r}$ and $r = \text{Pearson's } r$

1.65 = z score that defines change scores observed in 90% of cases

Equation 4 can be rewritten as

$$MDC_{90} = \sqrt{2} \times SEM \times 1.65 \quad \text{Equation 5}$$

For small sample sizes, the appropriate t multiplier replaces the z value at the 90% confidence interval (Equation 6)

$$MDC_{90} = \sqrt{2} \times SEM \times t \quad \text{Equation 6}$$

Where;

t = the appropriate t value at a $df = (n \text{ (number of pairs)} - 1)$, at an alpha level of 0.1

If a positive or negative change in scores greater than MDC_{90} occurs assessors can be confident that in 90% of cases the observed difference is not a chance finding.

A variation on this approach to describing reliability was proposed by Bland and Altman (1986). They proposed calculation of the mean difference between measures (d), the 95% confidence interval (CI) for d , the standard deviation of the differences ($SD \text{ diff}$), the 95% limits of agreement and a reliability coefficient. The Bland-Altman graph plots the difference between the measurements by two raters for each subject against the mean of the two measurements. The plot provides a visual representation of the level of agreement, assists identification of bias, outliers, and of an association between the variance in measures and the magnitude of a score. Hence if error was systematically greater for higher scoring students, this would be evident in a Bland-Altman plot.

The purpose of this reliability study was to determine the inter-rater reliability of physiotherapy educators in awarding clinical education grades to pre-entry level physiotherapy students using the APP. Both correlational coefficients and metricated errors were estimated to provide a comprehensive analysis of the likely utility of APP scores and to enable score and change score interpretation.

10.2 Method

The ideal approach to the study of reliability entails estimation of measurement error that can be expected during the typical application of the measurements. The method employed in this trial adhered to this principle as closely as possible within the constraints of the authentic clinical environment.

10.2.1 Study design

The inter-rater reliability trial was a cross-sectional study designed to replicate authentic measurement procedures. Two assessors (clinical educators) independently rated a student's level of professional competence using the APP at the end of a usual five week clinical placement block scheduled during one semester in 2008. Students provided supervised care/services to clients during this placement on a full-time basis (32-40 hours/week).

10.2.2 Recruitment of participants

Since not all physiotherapy education programs typically utilised shared supervision (i.e. two supervisors sharing supervision of a student), programs where this routinely occurred were identified from the twelve physiotherapy entry-level programs in Australia. Five universities were identified where this occurred (Curtin University, James Cook University, La Trobe University, The University of Sydney and Griffith University) and clinical educators were subsequently invited to participate in the trial (Appendix 10.1). To be eligible to participate, educators had to be working in pairs and each member of the pair had to be able to make sufficient observation of student performance to confidently complete the APP at the end of the five week placement. In addition, each participant had to be able to independently complete an APP assessment and remain blind to scores awarded by the partner educator.

Information on the reliability trial was provided in writing to the educators and students and their consent to participation was obtained. Assessment data were excluded from analysis if either the student or their clinical educator did not consent to participation in the research

or if any pair of assessors did not complete the APP instrument as per instructions. Participants were advised that all data would be permanently de-identified prior to data analysis.

No previous data were available with which to conduct power analysis regarding the numbers required to achieve significance for the obtained inter-rater score correlation. A minimum of 30 pairs of educators was set as the desirable recruitment target as this sample size typically produces data that conform to a normal distribution (Gravetter & Wallnau, 2005). In addition, if adequate evidence of reliability was not identified with this sample size, it would be unlikely that APP scores had properties required for confident interpretation of scores for an individual student.

10.2.3 Ethics approval

Ethics approval was obtained from the Human Ethics Committee of Griffith and Monash Universities and from the Human Ethics Committees of each of the participating universities (Appendix 3.4).

10.2.4 Trial preparation

10.2.3.1 Training of participants

All clinical educators received training through attendance at a three hour workshop and/or access to a clinical educator resource manual. The manual contained all information pertinent to standardised use of the APP instrument (refer to Appendix 9.1). As a component of all entry level programs, students were educated in the assessment process and use of the APP instrument using a standardised presentation developed by the research group. The training was conducted by a member of the research group or the clinical education manager at each university. Prior to commencement of the clinical unit, each participating clinical educator received all relevant documentation including a copy of the APP (version 4) instrument for each student, a clinical educator and student demographic data form and a feedback questionnaire (Appendix 6.5). A reply paid envelope was also provided to facilitate return of completed forms to the research team.

10.2.5 Trial procedure – during the clinical unit

Clinical educators were advised to conduct the clinical unit according to normal procedure. This meant providing formative feedback to students on their progress midway through the unit and summative assessment on completion of the five week unit. Thus the student was assessed by clinical educators who had numerous opportunities to observe the student's performance across multiple activities related to service delivery. The clinical units represented the major areas of physiotherapy practice and included musculoskeletal, cardiorespiratory, neurological, paediatric and gerontological physiotherapy. During the clinical unit the educators were instructed they could have normal discussions with colleagues about strategies to guide the student and facilitate learning but were requested not to discuss intended or actual grading of the student's performance. Educators were able to make contact with the researcher to clarify issues associated with participation in the study.

10.2.6 Trial procedure – on completion of the clinical unit

On completion of a five week clinical unit, the pairs of educators were instructed to complete the APP independently and seal the completed instrument in an envelope for return to the researchers prior to communication about assessment outcomes with the other clinical educator. Where the APP was the sole assessment instrument, educators were instructed to complete the APP independently and seal the completed instrument in an envelope for return to the researchers. Educators could then meet to discuss the student's performance and jointly complete an APP for return to the student's University. The student viewed this APP assessment prior to its submission to the university.

For the programs using the APP in parallel with a current university-specific form, the educators were instructed to assess the student's performance using the APP at the end of the clinical unit prior to completing the required university assessment documents. For these universities, students were informed that the scores provided by the clinical educators on the APP instrument would not be used in establishing their grade for the clinical unit and students did not view the completed APP forms.

10.2.7 Data management and analysis

Completed forms were returned to the researcher by mail. Item scores, total scores and Global Rating scores were entered into a spreadsheet, matched to the paired report and de-identified prior to commencing analysis by the research group (MDal, JK and MDav). Data analyses were performed using SPSS 18.0 (SPSS Inc.).

Planned data analysis included the following: a summary of descriptive statistics; comparison of mean scores for the first and second test result using paired-samples t-tests to assess for any unanticipated systematic differences between assessors (J. Pallant, 2005); calculation of Pearson's r and the Intraclass Correlation Coefficient (ICC 2,1) (two-way random-effects model) (and their confidence intervals), the SEM, the MDC_{90} , a Bland and Altman analysis for total and individual item scores, and a plot of the mean of scores for the two raters against the difference between the rater scores (Bland & Altman, 1986) to examine consistency in error across the spectrum of obtained scores. In addition, percentage agreement for decisions across raters in total scores, item scores and GRS scores was calculated.

10.3 Results

10.3.1 Participant characteristics

Thirty-three pairs of clinical educators (66 independent educators) and 33 third and fourth year physiotherapy students from five universities consented to participate in the reliability trial. Every data point was independent. No single pair of educators assessed more than one student which reduced the potential for enhanced reliability that may occur if educators regularly supervised students together. Three pairs were subsequently excluded as the educators completed the APP instrument a week apart, allowing for errors due to real changes in student performance over that time. Of the 60 clinical educators, 40 had participated in one of the field tests. Of the five Universities participating in the inter-rater reliability trial, two used the APP in parallel with their current university specific clinical assessment form and three used the APP as the sole assessment instrument. Participant details are presented in Table 10.1.

Table 10.1: Demographics for participants in the inter-rater reliability trial

Demographic	Curtin University	James Cook University	La Trobe University	University of Sydney	Griffith University
Programme & Year of study	3 rd Year (4-year bachelor degree)	3 rd Year (4-year bachelor degree)	3 rd /4 th Year (4-year bachelor degree)	3 rd Year (4-year bachelor degree)	5 th year (5-year double degree)
No. of students (M) (F)	1M, 3F	3M, 3F	2M, 4F	3M, 2F	3M, 6F
Student average age (years)	22.5(3.4)	22.3(3.5)	22.5(3.3)	22.6(3.2)	23.0(3.1)
No. of CEs (M) (F)	3M, 5F	4M, 8F	5M, 7F	4M, 6F	6M, 12F
CE average age (years) (sd)	39.5 (8.9)	36.5 (8.3)	33.3 (8.7)	36.4 (8.9)	35.4(8.6)
Facility type & clinical area/s	Hospital, outpatient musculoskeletal	Hospital, cardiorespiratory, paediatrics	Hospital, neurological rehabilitation Community Health centre, community health	Hospital, cardiorespiratory, gerontology rehabilitation.	Hospital, inpatient orthopaedics, outpatient musculoskeletal, paediatrics
APP (sole or in parallel)	Parallel	Sole	Sole	Parallel	Sole

Abbreviations: M=male, F=female, U/G=undergraduate, CE=clinical educator

10.3.2 Relationship between raters

10.3.2.1 Percentage agreement between raters

Ratings by two assessors for 14 of the 20 APP items were identical on 70% or more occasions (Table 10.2). Item 20 (risk management) demonstrated the highest (83%) percentage agreement and for item 19 (evidence based practice) agreement was lowest at 56.7%. All raters were within one point of agreement on the 5 point rating scale.

There was complete agreement between 24 pairs of raters (80%) for the overall global rating of student performance. The remaining six pairs of raters all scored within one point of each other on the 5 point global rating scale.

Table 10.2: Percent agreement between raters on item and global rating scores

Item	Absolute agreement	
	/30 pairs	%
1	24	80.0
2	21	70.0
3	22	73.3
4	18	60.0
5	23	76.7
6	21	70.0
7	24	80.0
8	18	60.0
9	21	70.0
10	21	70.0
11	22	73.3
12	21	70.0
13	21	70.0
14	23	76.7
15	21	70.0
16	21	70.0
17	20	66.7
18	18	60.0
19	17	56.7
20	25	83.3
GRS	24	80.0

10.3.2.2 Pearson's product-moment correlation coefficient

A scatterplot was visually assessed for violation in the data of assumptions of linearity and homoscedasticity. Figure 10.1 shows a positive, strong linear relationship between rater 1 and rater 2 for the total APP score (scale width 0-80). Based on definitions provided by Cohen (1988), there was a strong, positive and significant correlation between scores for the two raters [$r= 0.92$ (95%CI 0.87 – 0.95), $df = 29$, $p < .0005$]. The coefficient of variation implies that 85% (95%CI 75-90%) of the variance in a second rater's scores was explained by variance in the first rater's scores.

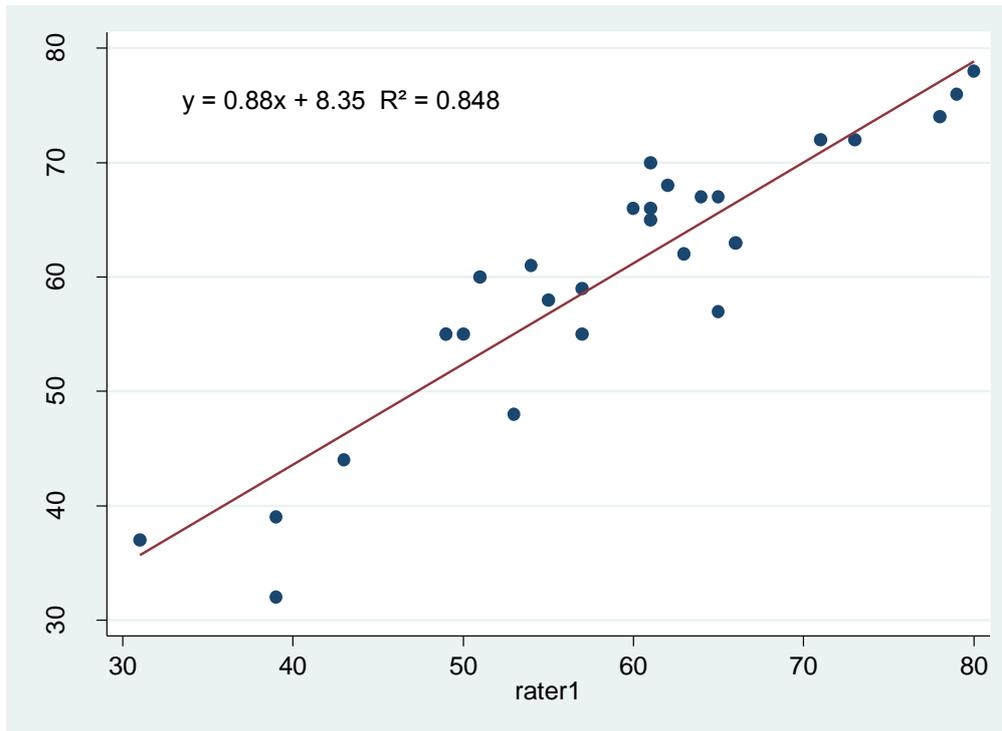


Figure 10.1: Scatterplot of APP scores for rater 1 and rater 2

10.3.3 Paired t-test

As expected, there was no statistically significant difference in the total APP score comparing rater 1 ($M=58.8$, $SD=11.83$) and rater 2 ($M=60.1$, $SD=11.34$, $t(29) = -1.49$, $p = 0.147$), nor was there a difference for the global rating scale (GRS) scores (rater 1 ($M=2.43$, $SD=0.77$), rater 2 ($M=2.5$, $SD=0.82$), $t(29) = -.57$, $p=0.573$).

10.3.4 Intraclass correlation coefficient ICC(2,1)

The Intraclass Correlation Coefficient 2,1 (two-way random effects model) for total APPs scores for the two raters was 0.92 (95% CI 0.84 to 0.96). The ICC 2,1 for the global rating scale scores was 0.72 (95% CI 0.50 – 0.86). Table 10.3 presents the ICC (2,1) results for the total score, each of the 20 APP items and the GRS.

Table 10.3: ICC 2,1 for items, domains of practice, GRS and total score on APP

Item no.	Wording	ICC 2,1	95% CI	ICC Sig. $p <$	SEM 95%CI	MDC ₉₀
Total APP score		.92	.84-.96	.0005	6.5	7.86
Global Rating Scale		.72	.50-.86	.0005	.84	.98
Professional Behaviour (items 1 – 4)						
1	Demonstrates an understanding of patient/client rights and consent	.81	.64-.90	.0005	.31	.69
2	Demonstrates commitment to learning	.70	.46-.85	.0005	.35	.70
3	Demonstrates ethical, legal & culturally sensitive practice	.77	.57-.88	.0005	.35	.77
4	Demonstrates teamwork	.65	.37-.81	.0005	.45	.64
Communication (items 5 – 6)						
5	Communicates effectively and appropriately - Verbal/non-verbal	.82	.66-.91	.0005	.30	.85
6	Demonstrates clear and accurate documentation	.79	.56-.89	.0005	.31	.80
Assessment (items 7 – 9)						
7	Conducts an appropriate patient/client interview	.80	.62-.90	.0005	.30	.80
8	Selects and measures relevant health indicators and outcomes	.60	.29-.77	.0005	.43	.61
9	Performs appropriate physical assessment procedures	.71	.48-.85	.0005	.38	.71
Analysis and Planning (items 10 – 13)						
10	Appropriately interprets assessment findings	.63	.35-.80	.0005	.37	.65
11	Identifies and prioritises patient's/client's problems	.75	.53-.87	.0005	.36	.74
12	Sets realistic short and long term goals with the patient/client	.76	.55-.87	.0005	.35	.75
13	Selects appropriate intervention in collaboration with patient/client	.73	.50-.86	.0005	.35	.73
Intervention (items 14 – 18)						
14	Performs interventions appropriately	.82	.66-.91	.0005	.29	.85
15	Is an effective educator	.82	.65-.90	.0005	.35	.81
16	Monitors the effect of intervention	.60	.32-.79	.0005	.38	.60
17	Progresses intervention appropriately	.76	.57-.88	.0005	.36	.77
18	Undertakes discharge planning	.71	.49-.85	.0005	.44	.71
Evidence Based Practice (item 19)						
19	Applies evidence based practice in patient care	.70	.43-.83	.0005	.44	.68
Risk Management (item 20)						
20	Identifies adverse events/near misses and minimises risk associated with assessment and interventions	.74	.52-.86	.0005	.34	.75

10.3.5 Standard Error of Measurement (SEM)

In this trial the SEM was determined for the total and individual item scores (see Table 10.3). The standard error of measurement for the total score was 3.2 APP points (scale width 0 – 80) indicating that a student’s true score will typically fall between an obtained score plus or minus 3.2 (at 68% confidence).

The 95% confidence band around a single score was calculated using Equation 7:

$$SEM \times 2.045 \qquad \text{Equation 7}$$

The 95% confidence band around a single score was 6.5 APP points (given $t(0.05, df= 29) = 2.045$). This implies that in 95% of cases a student’s true APP total score will fall between the obtained score plus or minus 6.5 points.

The SEM on the four point GRS was 0.41 indicating that a student’s GRS score was accurate to within plus or minus 0.41 points (at 68% confidence) or 0.84 (95% confidence) (Table 10.4). This implies that in 95% of cases a student’s true score on the 5 level GRS will fall between the obtained score plus or minus 1 point.

10.3.6 Minimal Detectable Change (MDC)

Minimal detectable change scores were calculated for the total and individual item score data at the 90% confidence interval. The MDC_{90} for the APP total scores was 7.86 (given $t(0.1, df= 29) = 1.699$). This implies that a change in score of around 8 APP total score units is required to be confident that for 90% of students demonstrating changes of this magnitude, real change in professional competence has occurred. As the APP scale width is 0 – 80, the MDC_{90} represents 9% of the scale. For each item the MDC_{90} ranges from 0.60 – 0.85. Therefore on the 5 point rating scale used to score each item, a change in rating of around 1 point (the minimal observable change) is required to be confident that real change in performance on that item has occurred.

10.3.7 Bland Altman analyses

A Bland and Altman plot was constructed to display errors in estimates of total APP scores (Figure 10.2). In this plot, differences between raters’ marks were plotted against the mean

of the two raters' marks and the 95% limits of agreement were defined. Data are evenly distributed above and below the y-axis, indicating no important systematic differences between raters. Errors also appear similar regardless of the magnitude of averaged scores, indicating that it is valid to apply a single error estimate in the interpretation of scores across the width of the scale.

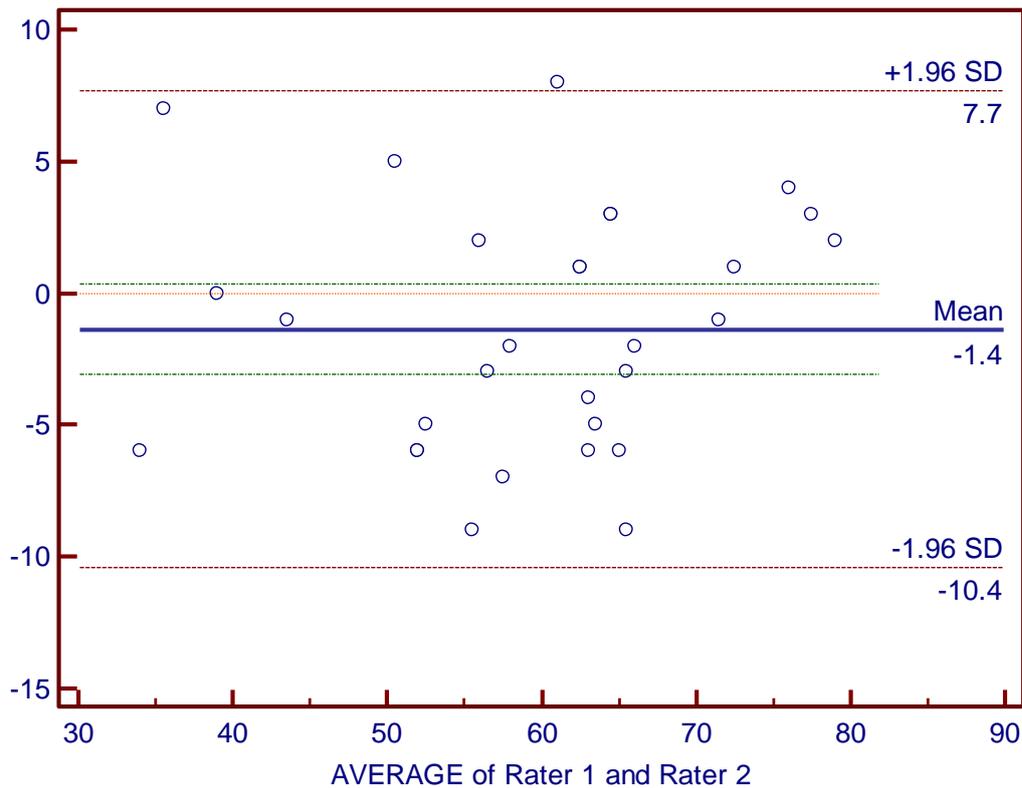


Figure 10.2: Plot of the differences between raters' marks against the means of raters' marks for the total score out of 80 (n=60 assessments). The x-axis bisects the y-axis at the mean difference between raters and the upper and lower lines represent the 95% limits of agreement.

Table 10.4: Results of previous inter rater reliability trials conducted in clinical environment relocate to after discussion commences

Author	No. of raters	No. of students	Blinding of raters	Mean test 1	Mean test 2	Test 1-2 mean difference(d) & SD of difference	Test1/test2 correlation
Taskforce 2002	2 x inter-rater reliability trials. Trial 1:70 pairs; Trial 2: 35 pairs	X	Yes	X	X	X	ICC 2,1, correlation range Trial 1: -.02- .62 (for each item) Trial 2: -.21- .76 (for each item) &.87 for total CPI score.
Coote	43 paired data sets. No. of raters unknown	X	Yes	X	X	Total score: Mean difference= 0.64, sd=5.35 MDC90 = 8.9	Total score: ICC= 0.84 95% CI .72 to .91 Assessment ICC=0.66 Treatment ICC=0.78 Professionalism ICC=0.78 Documentation ICC=0.82 Communication ICC=0.73 ICC: Total score= 0.62 (3 rd yr) - 0.59 (4 th yr). The 7 major competency areas showed ICC range 0.09 - 0.60. Professional behaviour had lowest ICCs 0.09 to 0.25.
Loomis	X	48 3 rd yr n=23, 4 th yr n=25	Yes	0 – 4 scale Mean score 4 th yr = 3.22, 3 rd yr = 2.72	Mean score 4 th yr = 3.19, 3 rd yr = 2.79	X	
Meldrum 2008	58	86 paired assessments. Exact number of students unknown.	Yes	X	X	1. Overall mark (out of 100) (d)= -0.5 (sd 2.9, 95%CI of difference: - 1.1 to 0.2, SE of diff 0.3,) MDC90=4.8 2. Patient management (out of 600) (d)=-2.3 (sd 19.8, 95%CI of difference :-6.5 to 2.0, SE 2.1) 3. Professional development (out of 300) , (d)= -0.9 (sd 12.4, 95%CI: -3.6 to 1.8, SE (1.3) 4. Organisation and management, n=38 (out of 100) 0.9 (sd 5.1, 95%CI - 0.8 to 2.6, SE 0.8)	ICC2,1 1. Overall mark (out of 100) 0.84 . 95% of the time markers will be within 6.2 marks of each other. 2. Patient management 0.75 3. Professional development 0.75 4. Organisation and management 0.81

X= Data not reported; **MDC90 calculated by author (MDal)**

10.4 Discussion

There is a limited evidence base investigating reliability of the marking of physiotherapy student performance during clinical placements (see Table 10.4) and no previous study has been conducted on the reliability of assessment of competency to practice physiotherapy in Australia.

In this inter-rater reliability study of APP scores, the percentage agreement for individual items and the GRS was high with 70% absolute agreement on 14 of the 20 items. Similarly there was complete agreement between raters for the overall global rating of student performance on 80% of occasions. Where there was a lack of agreement, all raters were within one point of agreement on both the 5 point item rating scale and the GRS.

As all raters were independent there was no expectation of systematic difference between raters. This was supported by the results of a paired-samples t-test which showed there was no significant difference in the total APP scores between the raters.

The Intraclass Correlation Coefficient 2,1 (two-way random effects model) for total APP scores for the two raters was 0.92 (95% CI 0.84 to 0.96). Individual item ICCs ranged from .60 for items 8 (selecting relevant health indicators and outcomes) and 16 (monitoring the effect of intervention) to .82 for items 5 (verbal communication), 14 (performing interventions) and 15 (being an effective educator). The ICC 2,1 for the global rating scale scores was 0.72 (95% CI 0.50 – 0.86). Four previous studies have investigated inter rater reliability related to the assessment of clinical performance of physiotherapy students and demonstrated similar results to this present study (Coote, et al., 2007; Loomis, 1985a; Meldrum, et al., 2008; Task Force for the Development of Student Clinical Performance Instruments, 2002). Intraclass correlations (2,1) of .87 for the total CPI score were found for joint evaluators of physiotherapy students and .77 for joint assessments of physiotherapy assistants (Task Force for the Development of Student Clinical Performance Instruments, 2002). Coote et al (2007) reported an ICC of 0.84, while Loomis (1985a) found ICC's of 0.62 and 0.59 for third and fourth year total scores respectively.

Caution needs to be applied when interpreting these reliability results because an ICC of .92 indicates that 85% of the variance in the second rater's scores are explained by variance in

the first rater's scores. The remaining 15% of variance remains unexplained error. It has been proposed that raters are the primary source of measurement error (Alexander, 1996; Landy & Farr, 1980). Other studies suggest that rater behavior may contribute less to error variance than other factors such as, student knowledge, tasks sampled and case specificity (Govaerts, et al., 2002; Keen, Klein, & Alexander, 2003; Shavelson, Gao, & Baxter, 1993).

It may be argued that an ICC = 0.92 reported in this study is high. While the paired assessors were advised to have no communication concerning marks or grading of student performance during the five week clinical placements, this may not have been achieved. Similarly, discussion between educators on strategies to facilitate learning in a student may have inadvertently communicated the level of ability being demonstrated by a student from one educator to the other. In defence of the results, in all thirty pairs of raters, education of students was shared with little, if any, overlap of work time between raters. This trial design limited opportunities for discussion between raters. The comprehensive nature of the training of raters in use of the APP instrument may have enabled informal norming to occur (a desirable outcome), positively influencing the level of agreement between raters. While the possibility of inadvertent communication between raters may be seen as a limitation of the inter rater reliability study, independent replication of the assessment process as it occurs in 'real life' was given priority and the possible limitations relating to this method were considered acceptable. Additionally as stated earlier, the results in this study are comparable to those of previous studies where the inter rater reliability trials were also conducted in the clinical environment during usual clinical placements (Coote, et al., 2007; Loomis, 1985a; Meldrum, et al., 2008; Task Force for the Development of Student Clinical Performance Instruments, 2002).

Although the ICC and SEM are related, they do not convey the same information. The ICC provides information on the level of agreement, whereas the SEM provides information on the magnitude of error expressed in the scale units of measurement. The SEM for the APP (3.2) represents 4% of the scale width. The reliability of the APP compares favourably with reliability estimates reported by others who have developed instruments for assessing competency to practice physiotherapy (Coote, et al., 2007; Meldrum, et al., 2008). Coote et al (2007) reported data that enabled calculation of the SEM and it appears that for their

instrument (Common Assessment Form) this was also 4% of a scale from 0-100. For Meldrum et al (2008) we calculated that error estimates were approximately 3% of a 0-100 scale. Hence we confer that clinicians are reasonably consistent in their judgements of student ability to practice and that this consistency is evident across different scales, countries and practice conditions.

The 95% confidence band around a single score for this data was 6.5 APP points. The 95%CI for interpreting individual items were also small, ranging from .29 for item 14 (performs interventions appropriately) to .45 for item 4 (demonstrates teamwork). Therefore, 95% of the time a student's true score on the 5 level item rating scale will fall between the obtained score plus or minus 1 point. Similarly the 95% CI for the error in the estimate of the four point GRS was 0.84. This implies that 95% of the time a student's true score on the 5 level GRS will fall between the obtained score plus or minus 1 point.

With a scale width of 0 – 80, an error margin of 6.5 (95%CI) is acceptable. This error enables a high level of accuracy in ranking student performance as evidenced by test/retest correlation of .92. In addition, in our data (see Chapter Eleven), students commencing workplace education typically obtain mean scores of approximately 45 APP points; by the end of their clinical training average scores are in the order of 60 APP points. Hence an error margin of 6.5 allows a clear view of average student progress across the workplace practice period. Seventy-seven percent of students change across the practice period by more than the MDC_{90} of 8 points. Of the 23% of students with scores that remain unchanged across 6 placement blocks, approximately 70% were relatively low performing students across all blocks while the others were consistently average (23%) to high (7%) performing students. The high retest correlations shown in this study provide evidence that educators using the APP are consistent in rating the relative ability of students. This is important for conferral of academic awards and for monitoring improvement in performance relative to peers.

The SEM 95%CI of 6.5 has implications for students whose score is within the borderline pass/fail range. If the pass mark is 40 out of the total 80 marks on the 20 items, then 40 minus 6.5 (33.5) might be considered an outright fail, while 40 plus 6.5 (46.5) might be considered an outright pass. The values in between would require a process for deciding on

further assessment for confidence that the student has an adequate level of professional competence on the items for which scores are poor. There are many possible sources of error in assessment scores and these are likely to be related to circumstances, educator, student and the interaction of these factors. If other indicators of student ability indicate competency, a mark as low as 34 may be acceptable. Alternatively, if other assessments indicate a student consistently performs in the borderline range, further practice and assessment (or tailored remediation) may be triggered even by grades as high as 47.

For the APP the magnitude of change in scores required to conclude that real change has occurred is in the order of 7.8 points which compares favourably to previous studies. Meldrum et al (2008) reported a MDC_{90} of 4.8 (0-100) and Coote et al (2007), 8.9 for the total scores on their respective instruments. There was insufficient data provided in reports of other inter-rater reliability studies to enable calculation of the MDC_{90} or standard error of measurement.

On the APP for each item the MDC_{90} ranged from 0.60 – 0.85. Therefore on the 5 point rating scale used to score each item, a change in rating of around 1 point (the minimal observable change) is required to be confident that real change in performance on that item has occurred. Similarly on the GRS the MDC_{90} was .98 which again means that a change of 1 point on the GRS is required to be confident real change has occurred.

Norman et al (2003) reported that for health related quality of life outcome measures, the change in measures of health outcomes that people typically consider to be important (minimal important difference) is approximately half a standard deviation of raw scores for a representative cohort. If we thought that the same 'rule' might be applied to interpretation of the APP, students achieving changes of half a standard deviation would change by approximately 6 APP points. Therefore scores changes in the order of 6-8 points are almost certainly going to be perceived by students and educators as an indication of progress.

Frequently measures of internal consistency (Cronbach's alpha) are reported as the reliability of a test. If this is the only statistic reported in relation to reliability, this value

should be interpreted cautiously (Streiner & Norman, 2003; Wilson, 2005). In Chapters 4, 7 and 8, Rasch analysis of APP field test data has demonstrated high levels of internal consistency with the APP exhibiting a person separation index (PSI) of 0.96. This result provides further evidence supporting an acceptable level of reliability of the APP when used to assess clinical performance of physiotherapy students.

Of the sixty raters, 66% had used the APP instrument previously during one of the field tests. The exact number of times each of these raters had used the instrument was unknown. This level of prior experience may mean that the results are not generalisable to novice users. Further research to investigate the impact the level of educator experience may have on assessing students using the APP is warranted.

When examination conditions can be standardized (e.g. using OSCEs) variability in the conditions of examination can be controlled, limiting the influence that these variations might play in the outcomes of assessment. When assessment of performance takes place at the health delivery interface, many factors combine to influence student performance and subsequent assessment outcomes. Some of these include patient, student and educator emotional states and behaviours, the complexities of individual patient circumstances and health needs, and students' past experiences with the level of challenge confronted under assessment procedures. Assessors may also be subject to halo effects and racial and sex bias (Hammond, 2009; Herbers, et al., 1989; Rowland-Morrin, Burchard, Garb, & Coe, 1991; Wang-Cheng, Fulkerson, Barnas, & Lawrence, 1995).

These conditions are likely to decrease the reliability of 'one-off' assessments. Conversely, a student assessed longitudinally across a range of circumstances and by a number of assessors has repeated opportunities to demonstrate both ability and growth in ability. In these conditions, assessment outcomes might intuitively be considered to more reliably reflect true ability, as the averaging of repeated measurements typically narrows error bands for measurements taken under highly controlled experimental conditions. However, determining reliability of assessments under circumstances when student ability is expected to change presents an added challenge. With adequate funding, student performance could be concurrently monitored and assessed by more than one educator and assessment

procedures refined until adequate concordance in grading is achieved. However, even if acceptable error is identified under such an approach, it is likely that, occasionally, unacceptable variation in assessment will still occur across individual assessors. A pragmatic, and perhaps less costly approach to optimizing reliability is to implement effective education in best practice in student assessment and strategies for developing a shared vision of expectations of performance (Epstein, 2007; Wass, et al., 2001b; Wilson, 2005). If this results in graduates who are typically considered competent, the profession might infer that assessment procedures had adequate reliability.

There will always be some lack of agreement between raters and defining the limits of tolerable disagreement is challenging. Some variability would be expected due to the unpredictable challenges of a complex health services environment combined with variable opportunities for educators to observe student ability across the spectrum of clinical skills. Despite these challenges, in this inter rater reliability trial the physiotherapy clinical educators demonstrated a high level of reliability in the assessment and marking of physiotherapy students' performance on clinical placements when using the APP. This was found despite the variability anticipated due to different areas of practice, types of facilities and a spectrum of educator experience.

11. Chapter Eleven: Validity, future research directions and summary

11.1. Introduction

Messick (1995b) defined validity as follows:

“Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment” (p471).

As with reliability, validity is not considered to be the property of an assessment instrument but of the meaning of the test scores that are generated through a combination of the assessment items, individuals completing the assessment and the context of the assessment. The emphasis is on how the assessment scores are applied and whether the score interpretation and use prevail across all people or population groups and all settings or contexts (Messick, 1996; Wilson, 2005).

As discussed in Chapter One, section 1.2, there are many approaches to establishing validity of assessment methods (American Educational Research Association, 1999; Baartman, et al., 2006, 2007; Fitzpatrick, et al., 1998; Kane, 1992; Kane, Crooks, & Cohen, 1999; Messick, 1989, 1994, 1995b; Streiner & Norman, 2003; Wilson, 2005; Wolfe & Smith, 2007a). The Standards for Educational and Psychological Testing (1999) advise that evidence of validity be assembled from multiple sources to substantiate the planned interpretations of instrument scores.

This thesis has outlined the rationale for and process of developing an assessment of physiotherapy students’ professional competence in the workplace, as well as the procedures undertaken to collect evidence regarding the validity of APP scores. The key question is whether the evidence supports the assertion that the assessment instrument can be validly used for the purpose for which it was designed. Using the five sources of validity evidence presented in the Standards for Educational and Psychological Testing

(1999), Chapter Eleven will present and discuss the evidence for the validity of APP assessment scores.

11.2 Sources of validity evidence: framework for instrument development

The overarching framework or model used in the development of an instrument provides the first and most important piece of the evidential argument for validity of measurements (Wolfe & Smith, 2007a). As presented in Chapters 2 and 3 in this research, the four building blocks approach proposed by Wilson (2005) was employed. This model includes definition and mapping of the construct of interest, design and compilation of the items which comprise the instrument, a strategy for coding the responses into an outcome space and scoring them and the fourth building block, the measurement or statistical model enabling interval level calibration of the construct through the scores. In this chapter various types of evidence that result from using this model and how that evidence (qualitative and quantitative) can be used to support validity arguments will be assembled. To assist the reader and to avoid unnecessary repetition of information already discussed during the relevant chapters, a summary of sources of validity evidence and how validity is supported by methods used in the development and evaluation of the APP is provided in Table 11.1. Additional validation evidence not presented in previous Chapters will be described in detail.

Table 11.1: Sources of validity evidence

Category of Validity evidence	Recommended types of possible validation activities	What was done in the development and evaluation of the APP
Test content:	<ol style="list-style-type: none"> 1. Construct Mapping: define instrument purpose 2. Logical analyses of extent to which the test content represents the content domain <ul style="list-style-type: none"> • Engage with experts in the field & all relevant stakeholders • Amplify quality of content through independent review of items • Test blueprint: map against relevant professional standards 3. Address item technical quality, 4. Develop scale and scoring methods 	<ol style="list-style-type: none"> 1. Construct Map <ul style="list-style-type: none"> • Discussion with purposively sampled experts including research team • Construct map drawn • Presented to stakeholders for feedback and modification 2. Item Design/Development: <ul style="list-style-type: none"> • Test content reviewed by panel of experts consisting of academics, clinical supervisors and clinical managers to address adequate representation of the construct. • Collated items from all existing instruments, relevant documents, theory and research • Assembled larger item pool than required • Engaged with experts in the field & all relevant stakeholders • Comprehensive cross section of users engaged in development • Standardised criteria for item reduction defined: technical as well as content • Independent item reduction by blinded reviewers • Engaged in independent review processes of items eg focus groups etc and pilot trial/field tests • Mapped content to relevant standards • Pilot and field testing phases – item content and design tested and refined via pilot and field tests. 3. Criteria applied to address item quality: <ul style="list-style-type: none"> • Target one attribute (explicit learning outcome); • Describe an observable and measurable behaviour; • Unambiguous, clear and defensible; • Important to students, educators and/or key stakeholders; • Culturally sensitive • Described without jargon; • Without value-laden words; • Concise; • Free of negative wording e.g. not or never

Response processes:	<ol style="list-style-type: none"> 1. Study how judges/observers collect, record, and interpret data. 2. Monitor changes in, or development of, responses to performance tasks eg., evidence of developmental progression in competency 3. Conduct rater training 4. Investigate test scoring eg., acceptable person fit statistics 5. Conduct student training: familiarity with format, purpose and score interpretation 6. Assess for theoretically congruent item hierarchies 	<p>4. Scoring system</p> <ul style="list-style-type: none"> • Reviewed purpose and context of the assessment • Reviewed all available applicable scoring systems • Decided on norm vs criterion referencing • Decided on level of measurement (categorical or continuous) • Decided on appropriate scale format • Addressed quality through independent review processes of scoring system eg focus groups etc and pilot trial/field tests • Developed appropriate scoring guides (eg training manual). • Provided examples of behaviours representative of the level of performance being measured (performance indicators)
Internal structure:	<ol style="list-style-type: none"> 1. Undertake Factor analysis 2. Conduct item analyses to examine item interrelationships. 3. Conduct differential item functioning (DIF) studies. 4. Investigate reliability: test/retest, inter rater, should include SEM 5. Investigate discriminatory ability of rating scale 6. Identify gaps in item content, and/or redundant items 7. Investigate dimensionality of instrument 8. Investigate targeting: floor or ceiling effects 9. Investigate sample and item independence 	<ol style="list-style-type: none"> 1. Field testing phase: <ul style="list-style-type: none"> • Conducted think aloud interviews • Conducted focus groups to gather information on: clinical educators and students • Provided questionnaires to clinical educators and students 2. One-way repeated measures ANOVA provided evidence of expected developmental progression in competency across time captured in APP scores. 3. Conducted CE training as per usual University training All CEs and students received training: resource manual and/or face to face workshops 3 hours focusing solely on use of the APP; 4. Conducted Rasch analysis to examine person fit, item difficulty 5. All students received training in all aspects of APP and score interpretation 6. Investigated item hierarchies via Rasch analysis <ol style="list-style-type: none"> 1. Conducted initial factor analysis 2. Data consistently displayed fit to the Rasch model 3. No consistent Differential Item Functioning observed (one instance of DIF for student gender on item 6, not confirmed in follow up study) 4. Measurements were highly consistent in reliability study SEM acceptable; error small relative to expected changes in scores 5. No floor or ceiling effects observed

- 10 Choose statistical measurement model that provides evidence of internal structure eg., IRT & RMM
 11. Apply generalizability theory

Relation to other variables:	<ol style="list-style-type: none"> 1. Conduct correlational studies between scores and external variables: convergent and discriminant 2. Conduct correlational studies of the extent to which scores forecast performance or scores obtained at a later date: predictive 3. Conduct known-group comparison studies, intended to test hypotheses about expected differences in test scores across specific groups of examinees. 	<ol style="list-style-type: none"> 1. Conducted correlational studies (convergent and discriminant evidence): There was a weak but significant positive correlation between the results of the clinical unit and the practical skill examination [$r = .31, n=94, p < .002$]. There was no correlation between either of the written assessment tasks (written exam and radiology assignment) and the practically focussed assessment items. There was also a weak but significant positive correlation between the two written assessment items [$r = .33, n=94, p < .001$]. 2. No investigation of predictive evidence of test score validity 3. Conducted correlational study across time for APP scores: Student total APP scores increased significantly with increasing hours of clinical experience [Wilks' Lambda = .28, $F(5,52) = 25.75, p < .0005$, multivariate partial eta squared = .71]
Consequences of testing/Educational Impact.	<ol style="list-style-type: none"> 1. Investigate impact on student learning: benefits and negative consequences 2. Describe method of determining pass/fail score; 3. Describe estimation of pass/fail decision reliability including estimation of the standard error of measurement at the cut score. 	<ol style="list-style-type: none"> 1. Conducted student surveys and focus groups 2. Method and reasoning for determining pass / fail score described: entry-level professional competence 3. SEM 95%CI for score calculated 4. Performance indicators constructed to assist student learning by clearly stating expected behaviours in relation to each item
Additional criteria in support of argument for instrument validity (Bartman, et al., 2006)		
Authenticity Acceptability Costs	<ol style="list-style-type: none"> 1. Authenticity: assessment process resembles the criterion situation 2. Acceptability: assessment process is acceptable to stakeholders, includes time to complete and user satisfaction 3. Costs of implementation and ongoing use of an assessment approach 	<ol style="list-style-type: none"> 1. All testing conducted in authentic clinical environment 2. Acceptability canvassed via focus groups and survey from both CEs and students 3. Other than staff time associated with completion of the APP instrument , costs were not investigated

11.3 Validity evidence based on content

As recommended by the Standards for Educational and Psychological Testing (1999), during the development and evaluation of the APP instrument, evidence based on test content included empirical analyses of the adequacy with which the instrument content represented the domain of physiotherapy professional competence and of the relevance of the content domain to the proposed interpretation of the instrument scores.

Mapping of the instrument content against relevant professional standards also enabled blueprinting of the items against the construct ensuring adequate coverage of the domains of practice anticipated to be examined through assessment of clinical performance. The Australian Physiotherapy Council Standards (2006) guided the initial development of the item content and a formal mapping exercise was conducted on completion of the first field test once refinements to the instrument had been made.

The transparent consensus approach to instrument development, with input from an appropriate spectrum of key stakeholders (e.g. national accrediting bodies, graduate employers, educators who use the assessment procedures, academics, students and a broad spectrum of practitioners) provided evidence for validity based on test content. Experts' evaluations of the sufficiency, relevance and clarity of the components of the instrument were integral to provision of validity evidence relating to content (Cook & Beckman, 2006; Goodwin, 2002). Additionally educators reported that the inclusion of performance indicators (behavioural descriptors) for each item provided information the educator could use to support the process of judgement of student performance.

Stakeholder input via focus groups, interviews, workshops, surveys and teleconferences occurred at all stages throughout the instrument development and testing, enabling continual refinement of item content based on both qualitative and quantitative data collected across time through three field tests in the authentic clinical environment.

Evidence of the relevance and comprehensiveness of content coverage in the items and their associated performance indicators is shown by the final qualitative data collected during Field Test Two, where no further modifications or additions to these were requested by any of the stakeholders. In addition the positive feedback from educators and students regarding the usefulness of the performance indicators in understanding the items and scoring them was also evidence of validity of the instrument content.

Documentation of the rigorous steps taken in development of the APP, as detailed in Chapters 2 and 3, is a necessary first step in establishing a chain of validity evidence to support applications and interpretations of a measurement instrument. It also lays the foundation for the other aspects of validity. As Wilson (2005) states:

“Evidence based on content is essential because it contains the realisation of the construct, and that is what all the other aspects of validity play off.”p157.

11.4 Validity evidence based on internal structure

Internal structure, as a source of validity evidence, relates to the psychometric properties of the instrument and the measurement model used to score and scale the assessment (Downing, 2003; Terwee, et al., 2007).

In this research both factor and Rasch analysis enabled investigation of the internal structure of the newly developed APP instrument. Evidence of structural validity supported the interpretation that a student’s score on the APP is an indication of their underlying level of professional competence as demonstrated during clinical placements.

Data from both field tests demonstrated that the APP was robust when tested against the assumptions of the Rasch measurement model, with the independent t-test analysis supporting the assumption of unidimensionality. Factor analysis of Field Test One and two raw scores also supported the concept of unidimensionality of the APP, with one dominant factor explaining 59% and 61% of the variance respectively.

Issues of bias also pertain to the internal test structure category of validity evidence.

Differential item function (DIF) studies demonstrated there was no item bias in either field test for the following variables: student age, gender and level of clinical experience, clinical educator age, gender and experience as an educator, facility type, University attended and clinical area. This indicates the APP item ratings were not systematically affected by any of these nine variables. In addition in both field tests there appeared to be acceptable matching of item difficulty with person abilities and an even spread of items across the full range of student scores, suggesting effective targeting of the APP items.

Another validation activity within the internal structure category of evidence is the investigation of reliability (Cook & Beckman, 2006; Downing, 2003). In this research an inter rater reliability trial was conducted during normal clinical placements. This approach enabled independent replication of the entire measurement process as it occurred in real life (for full details refer to Chapter Ten). The physiotherapy clinical educators demonstrated a high level of reliability in the assessment and marking of physiotherapy students' performance on clinical placements when using a standardised instrument, the APP. This was evident across different clinical skill areas, types of facilities, and level of educator experience.

Overall the process of investigating the structural validity of the APP instrument and the strong reliabilities demonstrated, confirm Friedman and Mennin's (1991) suggestion that sampling many specific behaviours over time and in various situations, as occurred in this research during the clinical placements spanning 4-6 weeks, may provide an approximation of a student's true performance.

11.5 Validity based on relations to other variables

This source of evidence is primarily based on correlational studies and seeks to provide both confirmatory (convergent), discriminant and predictive evidence of test score validity (Downing, 2003; Streiner & Norman, 2003). Another important purpose of this external aspect of validity is to document, where relevant, anticipated between-group and within-person changes over time concerning the target construct (Wolfe & Smith, 2007b).

11.5.1 Convergent and discriminant evidence

Because there are no gold standard instruments for assessing professional competence in physiotherapy students, correlating data from a newly developed instrument and an existing assessment instrument is of unknown value. In this research the assessment instruments currently used in Australia and New Zealand lacked documented evidence of validity (Chapter One). Thus comparison of the newly developed APP with current instruments was not appropriate (Smith, 2001).

Evidence of validity of the APP instrument scores based on relations to other variables was investigated by examining the correlation of student scores on one assessment task hypothesised to measure a construct closely related to clinical performance and two further

assessment tasks considered to exhibit low correlations to performance in the clinical environment (refer to Chapter Eight for full details).

While the results of the Pearson product-moment correlation for the four different assessment tasks examining orthopaedic physiotherapy knowledge and practical skills lend support to the validity of the APP instrument based on relation to other variables, it is far from definitive. As Hobart et al (2007) state the limitation of this approach to establishing validity is that, to show that scores from an instrument do not correlate highly with measures of a dissimilar construct or correlate highly with measures of a similar construct, tells us nothing about what the actual scale measures. This approach informs us only that the two are related or not. However, as proposed in Chapter Nine, when this source of evidence is combined with evidence of face validity, alignment to required practice standards and widespread stakeholder support that the items are comprehensive, an argument is built that it is likely that the scores measure the construct that the profession calls 'entry-level professional competence'. Further research investigating convergent and discriminant evidence in relation to the APP results could provide a greater understanding of this aspect of instrument validity.

11.5.2 Predictive evidence

While this is an important aspect of the cumulative validity argument for the assessment approach using the APP instrument, the scope and design of this research did not allow for investigation of the aspect of validity related to predicting future performance.

Prediction of performance presents an even greater challenge when investigating instrument validity. Often performance based assessment scores are hypothesised to correlate with achievement of candidates on national licensure/certification examinations (Task Force for the Development of Student Clinical Performance Instruments, 2002). This may or may not be a reasonable hypothesis given the different knowledge, attitudes and skill sets required for a written examination compared with those required in a work-based placement.

11.5.3 Developmental Progression in Competency

Another source of evidence supporting the validity of the APP scores is the way in which increasing levels of ability, reflected in both Rasch and raw scores, related to increasing

levels of experience (time spent in clinical practice). Using a fixed reference point for assessor's judgements enabled demonstration of progress over time. For the APP, the fixed reference point represented the standard expected upon completion of training (entry level competency). It was hypothesised that experienced students in their final clinical placement (20-30 weeks) would perform better than novice students in their early clinical placements (0-9 weeks). The correlation of clinical blocks against APP scores showed a high correlation between time and score with predictable improvements as students gained experience [Wilks' Lambda= .28, $F(5,52) = 25.75$, $p < .0005$, multivariate partial eta squared = .71]. As competency should grow with experience, this observed correlation lends credence to APP scores as indicators of anticipated within-person changes over time relating to the target construct of professional competence. It is, however, important to note that no single test can unequivocally provide all validity evidence in this category. It is a continual process of testing, evaluation and refinement (Messick, 1996; Streiner & Norman, 2003).

11.6 Validity based on response processes

Validity evidence based on response processes relates to the substantive aspect of validity as described by Messick (1989). It is closely related to the previous section on content but focuses more on operationalisation by examining if assessors are using the instrument as was anticipated (American Educational Research Association, 1999; Messick, 1996). Studies that examine how data is collected, recorded and interpreted are recommended as a method to investigate this category of validity evidence (Goodwin, 2002). During this research validity evidence based on response processes was investigated through the use of think aloud interviews, and focus groups which examined the processes of the educators, how the scores were being interpreted and if there was consistency between this behaviour and what was intended (Downing, 2003) (Chapters 6 and 8).

11.6.1 Think aloud interviews and focus groups

The data obtained through interviews and focus groups revealed that overall the educators were completing the APP instrument as directed. There were some issues of misinterpretation and confusion initially relating to how to score items 6, 18 and 19, use of not applicable and the passing standard used when completing the GRS at mid unit. These issues were resolved through modifications made to the instrument and training materials.

Only item 6 continued to be scored unpredictably in the second field test reflected in both the findings of the interviews and focus groups and the results of Rasch analysis. These results from interviews and focus groups provide support for the validity of the APP based on response processes.

11.6.2 Rating scale analysis

For items scored on a polytomous scale, evidence of validity based on response processes can also be provided by investigation of the rating scale and how it functions (Wolfe & Smith, 2007b). In addition to use of the think aloud interviews to investigate this aspect of validity evidence, the Rasch model provides evidence relating to rating scale function. As discussed in Chapters 7 and 8, the two indices that are most informative include the ordering of thresholds and the person separation index (PSI). In both field tests Rasch analysis demonstrated a high PSI (0.92-0.96) and all twenty items exhibited ordered thresholds thus providing evidence of validity based on response process (for full details refer to Chapters 4, 7 and 8).

11.6.3 Rater and student training

If an instrument requires one person to rate the performance of another, evidence supporting response process could include evidence that educators (raters) and students have received some level of training (Cook & Beckman, 2006; Downing, 2003). Qualitative data collected via numerous methods during field testing (Chapters 6 and 8) revealed that all participants agreed on the importance of training clinical educators and students in correct use of the assessment instrument. Consistently themes relating to training resources included accessibility, standardisation, time efficiency, variety in modes of training delivery, clarity and succinctness of information, definitions of performance standards, role of formative and summative assessment, involvement of students in self assessment, exemplars of performance standards and guidelines for practise in use of the instrument. These issues were addressed when training resources were developed. Similarly, student familiarity with and understanding of instrument format, purpose and score interpretation provided evidence in this category and was also examined using focus groups and questionnaires. While training during the field tests was limited, it was similar to that usually provided to educators and students by each University. A level of standardised

training and regular review of the training materials and upgrading as appropriate is essential if correct use of the APP instrument is to be maintained and extended.

11.6.4 Item hierarchy

Another method of providing validity evidence based on response processes is to demonstrate that the empirical order of item difficulties agree with those predicted from the theory upon which the instrument was based (American Educational Research Association, 1999). Rasch analysis of data in the pilot trial and both field tests (Chapters 4,7 and 9) examined the sequence or hierarchy of average difficulty of the 20 items on the APP on five random samples. In each sample the item hierarchy of difficulty was similar with the first six items representing professional behaviour and communication the least difficult items, whereas the most difficult items related to analysis and planning, the application of evidence based practice to patient care and patient management. These item hierarchies of difficulty are consistent with theoretical propositions that students often experience greatest difficulty with items related to analysis and interpretation of clinical findings (clinical reasoning) and generally find professional behaviour items easier to achieve well in, even during their first clinical units (Ajjawi, Loftus, Schmidt, & Mamede, 2009; Kassirer, Wong, & Kopelman, 2010).

11.7 Validity based on the consequences of testing (educational impact)

Evidence of consequences is the most sporadically reported category of validity evidence and generally is the least reported evidence source when instruments used to assess clinical performance are investigated (Beckman, et al., 2005). This is despite it being described in the current Standards for Educational and Psychological Testing (1999). While it is not possible to anticipate all potential uses and misuses of test scores, this does not release the test developer from the responsibility of considering these aspects of test use (American Educational Research Association, 1999; Messick, 1995a). Because of the recent introduction of the idea of validity evidence based on consequences and the ongoing debate about whether it belongs in validation theory and practice, there are few documented methods to estimate this validity evidence. The Standards for Educational and Psychological Testing (1999) recommend that positive and negative (intended or unintended)

consequences of the testing process be examined and evaluated. This includes, for example, any change in student knowledge or skills, improved patient outcomes and student views on the instrument and its implementation. Provision of formative feedback using an assessment instrument and the impact of this on student learning, would also contribute to an understanding of the consequences of testing.

The passing score, the process used to determine the cut score, and the statistical properties of the passing score, for example, the standard error of measurement may also be included in this category of validity evidence (Prescott-Clements, et al., 2008)

11.7.1 Impact on student learning: benefits to learning and any unintended negative consequences

In the development of the APP focus groups with both educators and students were conducted throughout the different phases of the research and have provided insight into beneficial and potentially negative aspects of assessment of clinical performance (Chudowsky & Behuniak, 1998). The use of the assessment instrument to provide formative feedback midway through the clinical unit was well received by all stakeholders. In particular, both educators and students reported that the performance indicators were very useful especially as they were written as observable behaviours, assisting educators to give specific feedback to students on the areas of their performance that were adequate and those requiring improvement (refer to Chapters 6 and 8 for full details). Students reported that the items and performance indicators were easy to understand, comprehensive and promoted transparency of performance expectations between students and educators. These ideas were supported by the results of the educator and student based questionnaires.

The assessment process using the APP instrument requires the student to be rated by their clinical educator who has worked directly with them throughout the entire clinical unit. This provides the educator with numerous opportunities to observe the student's performance across multiple clients. Additionally, the design of the assessment is based on a formative assessment being carried out mid way through the clinical unit. This facilitates student learning by allowing the student to engage in a low risk, non-summative assessment of their performance that is focussed on learning and improving performance (Boud & Falchikov,

2006; Higgs & Bithell, 2001; McAllister, et al., 2010). Formative assessment also provides the student with the opportunity to identify if their perception of their performance is similar to that of their educator, assisting them to develop skills of self monitoring and reflection. Additionally, employing both formative and summative assessment requires the educator to provide specific examples of a student's performance to support their judgement. This assists the final judgement of the student's level of professional competence during summative assessment to be based on specific evidence of sufficient quality and quantity. Finally, if assessment drives learning then authentic work-based assessment is more likely to promote the type of learning required by students to become safe, effective health care practitioners which is the approach followed during the development and evaluation of the APP instrument (McAllister, et al., 2010).

While examination of any unintended negative consequences is difficult, the most potent evidence that unintended negative consequences were avoided is provided by the differential item functioning results found in both field tests and outlined previously in section 11.4. No significant DIF was demonstrated in either of the two samples in field test. While the aim of achieving totally scientific and objective assessment of human behaviour is unachievable, maintaining regular, comprehensive and standardised training of educators in the use of the APP instrument is the approach most likely to limit potential rater error and bias.

11.7.2 Method of determining passing score determining pass/fail score and estimation of the standard error of measurement.

Some university programs have traditionally used entry-level competencies as the benchmark against which to judge student performance, while others have used the performance that would be expected at the particular stage of the course (e.g., second year standard, third year standard). Despite initial concerns, the majority of educators supported the use of a fixed reference point for judgements. An advantage of marking students against acceptable entry level standards is that, theoretically at least, all assessors could assess against a set standard. In discussions about entry level/beginning physiotherapist standards there was clear consensus from participants that for consistent use of an instrument across

programs, students should be judged on each item against the minimum performance targets expected of a novice (entry-level) practitioner. Focus group participants agreed that many students had only one clinical unit within which to gain skills in core areas of practice e.g., neurological rehabilitation. It was therefore essential that the pass standard at the end of that block was entry level practice. The target of clinical education was the acquisition of a minimum acceptable level of skills irrespective of when each clinical unit was completed. The target of entry level competence enabled ranking of students relative to a common standard. In addition to presenting information on the process used to determine the passing standard, Norcini (2003) and Downing (2003) state that statistical properties of the scores also relate to the consequential aspect of validity. In this research calculation of the standard error of measurement demonstrated that the 95% confidence band around a single score for this data was 6.5 APP points. This implies that 95% of the time a student's true APP total score, will fall between the actual score plus or minus 6.5 points which is acceptable given the breadth of the score range from 0 to 80. These data provide further evidence of validity based on consequences.

Similar to the findings of Prescott-Clements et al (2008), the majority of educators indicated on the questionnaires that they were comfortable using both the five point item rating scale, the global rating scale and the performance indicators (84% of all educators from both field tests).

11.8 Additional sources of validity evidence

While not part of traditional validity frameworks, acceptability (which includes time to complete and user satisfaction), and costs of administering performance based assessments are important features of the complex assessment process necessary to appropriately assess professional competence. Investigation of these aspects provides relevant information as assessment processes are not only influenced by educational factors, but also by financial, managerial and institutional values (Baartman, et al., 2007; Streiner & Norman, 2003). Costs can be defined not only in monetary terms but also in terms of staff time. If the costs of implementation and ongoing use of an assessment approach in the clinical context are too high the approach is likely to be conducted with insufficient attention given to quality. Assessors and learners should find the assessment approach

manageable within the constraints of a busy clinical environment. The time taken to complete the APP was considered acceptable in both field tests. Despite the time taken to complete the assessment process ranging from 8 to 120 minutes (mean (SD) completion time FT1 =21.65 (13.3) mins; FT2 = 29.04(19.3) mins), the majority of educators reported that the APP was practical and easy to use in the clinical environment (83%). These time frames for completion compare favourably to the time taken to complete similar instruments with the CIET taking 30-60 minutes and the Blue MACS 1.6 hours (Chapter One). Comprehensive investigation of all costs incurred in association with assessment of professional competence of physiotherapy students in the clinical context was beyond the scope of this research. This important area of cost-efficiency warrants further investigation in future studies of clinical education. Nevertheless the instrument and the training manual are freely available to all, so there are no financial barriers associated with accessing the instrument.

11.9 Future research directions

The APP provides a sensible method for assessing competence to practice when used to assess students in workplace practice. It is likely that the needs of health service providers will change in the future. The methods described in this thesis could be utilised to adapt the APP to meet those (as yet unknown) needs. The simplest adaptation would be the development of novel performance indicators. These provide a method for defining new learning targets without changing the items or scoring of the APP. It may also be of value to develop a more expansive set of indicators for a range of typical practice environments. The APP could be modified more substantially if necessary through the addition of new domains, and this would be sensible if the domains of competency described by the accreditation body were revised in the future.

The APP has been developed to provide a level of certainty in assessment of clinical performance: the certainty that two different assessors would be likely to score a student in a similar way, the certainty that scores for the APP align with acceptable reference standards such as global rating scales, and other 'certainties' such as the likelihood that students who are strong will be identified as strong students across multiple workplaces by multiple examiners. This 'certainty', such as it is, sits within the complex teaching and

learning environment of the workplace practice. It does not counter the unpredictability of the challenges faced by students and educators, but its application in longitudinal assessment means that students have many opportunities to demonstrate their ability to operate intelligently within complexity. With only one assessment method, the challenge of supporting educators to gather maturity in student assessment presents as achievable; quality methods and examples, once developed, have widespread application; educators can talk to each other in an assessment language that is common, and through negotiation, learn how others think and act in response to challenging decisions.

This research found that the APP provides a valid person measure for each student on most occasions of its use as long as key aspects of its content, format, and procedure are adhered to, e.g. it cannot be assumed that the final assessment is valid if a mid assessment is not conducted as it may be that the mid assessment informs the final decision regarding competency. This aspect of instrument use will benefit from further research.

Delaney and Molloy (2009) argue that changing patient populations and changing practice knowledge driven by research and global changes in communication and information technology add to the complexity of the learning environment. Institutional and organisational factors related to the hospital, community and university, varying policy directives and agendas, limited resources and hierarchies of clinical decision making, all impact on teaching and learning in the workplace environment. These layers of complexity and the influence they have on the effectiveness of teaching and learning in clinical education, could overwhelm the educational researcher and limit rather than facilitate study into the unique teaching and learning opportunities that the clinical environment provides. This research attempted to examine the questions that educators and students have increasingly asked: “does the clinical assessment instrument really measure my level of competence?”, “is it reliable?”, “will the grading of my clinical performance be independent of the type of facility where I complete my clinical unit?”

Clinical educators and students who participated in the field tests requested video exemplars of students demonstrating varying levels of performance as a strategy that might assist in training of both students and educators. The most frequently requested exemplar was of a student performing at a passing standard. A DVD of exemplars of varying levels of student performance has been developed and is currently in use as a training resource. The DVD was produced following completion of the second field test. Further research into the

effectiveness of this additional training resource in promoting standardisation of assessment practices by educators and the impact on student learning is planned. It may be of value to convert the training resource manual into a self-study package for educators. This would provide an optional pathway for skill acquisition in using the APP. Production of this resource has commenced, and tests of its effectiveness are planned. There are many reasons why conversion of the APP to a web-based assessment instrument might benefit the profession. Centralised data collection, with all the opportunities this might yield (such as streamlining data collection and benchmarking), is planned. The proposed research includes collaboration with communication sciences (speech pathology) as this professional group have developed an on-line version of their profession specific assessment instrument COMPASS® (McAllister, et al., 2010). Opportunities for refinement of the instrument will regularly arise. Future research will springboard from these opportunities. It is important that the APP evolves across time to meet the changing needs of the profession and the health care system.

The APP has potential for application in the assessment of skills that are additional to those required for successful completion of Australian and New Zealand entry level training programs. There are a number of projects currently underway where the APP is being applied to the assessment of clinicians seeking specialisation qualifications, physiotherapists re-entering the profession and physiotherapists seeking registration in Australia. As the domains of competence in the APP are relevant to the assessment of specialised or advanced areas of practice, there are no obvious impediments to developing performance indicators that describe the advanced practice targets of postgraduate students. Data collected when the instrument was used in this way could be analysed for validity using methods described in this thesis. It would be informative to observe whether the hierarchy of difficulty for items changed with independent practice.

The impact of the assessment tool on learning was identified as a potential threat to consequential validity particularly given that competency based assessment has been criticised for negatively affecting learning (Wolfe & Smith, 2007b). Substantial effort was devoted to ensuring that the assessment tool had strong content validity such that the content directed students' attention, and that of their clinical educators, to appropriate learning goals. It was also proposed that this was important for valid engagement in the assessment process by students and clinical educators (Chapter 6), further safeguarding

consequential validity. A high degree of success was indicated by the strongly positive student and clinical educator's feedback regarding the content of the assessment (Chapters 6 and 8). In this research, evidence of validity related to the positive and negative consequences of testing (educational impact) was investigated. Quantitative evidence supported a decision that unintended negative consequences associated with inconsistent use of the APP were avoided. Differential item functioning analysis demonstrated no systematic differences in APP use associated with student age and clinical experience, clinical educator gender, age and experience as an educator, University, or clinical area. Individual variation in the use of the instrument does occur, and may be best countered by effective and regular training in standardised use of the instrument (Bursari, et al., 2006). Video scenarios that enable educators to rate students and compare and discuss peer ratings have been developed and are currently being tested. An electronic version of the APP that can be accessed and completed on line is under development. When this is completed, a near exhaustive list of performance indicators that explicitly describe a broader number of measurable behaviours can be developed with stakeholders. Additionally substantial benefits to teaching and learning may accrue through ongoing use of the features of Rasch analysis to evaluate the impact of different teaching and learning practices. For example future research using DIF analysis may yield useful information regarding the impact of different curricula upon the development of workplace competencies.

Student perception of their assessment using the APP, and their perception of the APP's effect on learning, was investigated using focus groups and student surveys. Overall consensus was reached by the student groups on the adequacy of the training they had received prior to commencement of the clinical unit and that the items and performance indicators were easy to understand, comprehensive and helpful in directing their learning. Student responses regarding feedback received during the clinical unit were variable. Overall however, the responses demonstrated general agreement that feedback was specific, given in a timely manner, usually based on the performance indicators and helpful in assisting them to improve their performance. The majority of students agreed that their clinical educators were well prepared in relation to use of the APP and that the marks they received were a fair indication of their performance. A project is planned for commencement in 2012 where one on one interviews of final year students will be

conducted to further explore the perceived effects of the assessment process on the nature of learning during clinical units.

The use of the assessment instrument to provide formative feedback midway through the clinical unit was considered valuable by all stakeholders. Both educators and students reported that the performance indicators were very useful especially as they were written as observable behaviours, assisting educators to give specific feedback to students on the areas of their performance that were adequate and those requiring improvement (Chapters 6 & 8). Educators reported that the performance indicators helped them to unravel some of the more complex aspects of clinical performance they identified as difficult to explain precisely to students. As reported by Molloy (2006), breaking down broad domains of practice into specific observable behaviours can assist in making the 'hidden curriculum' more explicit and achievable for students. Students also reported that the items and performance indicators promoted transparency of performance expectations between students and educators. Additional research into utilisation of formative feedback using the instrument to enhance student learning is planned. Questions that will be addressed include the role of formative peer assessment, the effects of feedback that is provided more frequently than at the halfway mark, and the effects of feedback that is supported by structured exercises developed to address specific learning needs.

In the thesis sociocultural theories were used to frame the way students learn, which means acknowledging that the learner is dealing with complexity, uncertainty, and continual changes in service provision ethos and practice (Rogoff, 1990; Vygotsky, 1986). On the basis of this layered and more nuanced recognition, it was argued that the best way to gather a reliable and valid representation of students' skills in clinical practice was via longitudinal monitoring of students' performance. Such longitudinal assessment encouraged observation of practice in a range of learning circumstances. In this way, assessment was viewed as an opportunity for educators to provide learners with clear, practical and relevant information and direction, and to help the learner develop skills of self-evaluation and self-regulation. The APP has been presented as one practical example of developing clear criteria linked to explicit and detailed performance indicators. These performance indicators were developed to assist in reducing assessor bias and to provide students with clear practice goals. Such detailed, and transparent expectations grounded in the realities of students' learning

experiences assists them to 'unpack' and make sense of their professional discourse and clinical practice. The assessment tool was also used to encourage students to reflect on their own performance in relation to the explicit behavioural descriptions.

Identifying competency to practice is important. It protects patients by demanding an acceptable level of knowledge and skills. It protects practitioners by demanding skills that enable them to sustain the delivery of services without placing themselves at risk of injury. It protects the professional reputation of practitioner by providing practitioners who can be trusted to work to the desirable practice standards. The use of the same assessment instrument across multiple higher education institutions assists in reducing irregular performance standards; it provides a consistent framework of expected student skills.

Aligned to the findings of Knight et al (2007) and Notzer and Abramovitz (2008) it became clear as this research was conducted that clinical educators need to be regularly mentored in effective formative and summative assessment. Systems need to be built to ensure that good teaching and learning resources are available and regularly utilised by educators, and that novice educators have time and opportunity to develop essential skills (Bursari, et al., 2006). Consistency in the use of the APP assessment instrument allows participating universities to collaborate in developing opportunities and resources for training of clinical educators. This is in contrast to the very difficult position faced by educators when there was in excess of 25 different instruments used Australia wide.

Safety and risk management are core aspects of the assessment process reflected by the inclusion of this domain of practice as an individual item (item 20). There was considerable stakeholder debate during item development regarding whether or not it should be compulsory to pass item 20 to achieve an overall pass for the practice unit. The consensus of stakeholders was that a student needs to learn about safety and risk management during a clinical unit in the same way they are required to learn about the other 19 items. If however during the summative assessment the educator considers the student's clinical practice to still be 'unsafe' then not only would the student fail item 20 but their unsafe approach to practice would be reflected in inadequate performance in several other domains of practice eg., assessment, planning and intervention. Thus the student exhibiting 'unsafe' practice would be likely to fail multiple items and be graded as unsatisfactory for the practice unit. This would usually lead to an opportunity to remediate these practices before progressing in the clinical program. The APP can also be used to provide students

who require remediation with specific learning targets for which further learning opportunities can be constructed.

11.10 Chapter Summary

This Chapter outlines the evidence of validity currently available for the newly developed APP instrument to assess physiotherapy student performance in the work place and recommends aspects requiring research in the future to more fully investigate this instrument, its use by educators in the clinical environment and the impact of the instrument on the nature and quality of learning relationships. Using the five sources of validity evidence presented in the Standards for Educational and Psychological Testing (1999), this Chapter has provided data supporting the assertion that the APP instrument can be validly used for the purpose for which it was designed, workplace-based assessment of professional competence of physiotherapy students. This research has therefore delivered an important outcome for physiotherapy education in that a single instrument with known validity and reliability is now available to replace the twenty-five distinct assessment forms formerly being used. To date (2010) 17 out of 18 Universities in Australia and New Zealand have adopted the APP as the sole assessment form, and a further three new programs commencing within the next two years are also adopting the instrument. This research will assist the physiotherapy profession by ensuring that graduate physiotherapists enter the workplace well equipped to provide quality care to their future clients, the ultimate goal of any professional preparation program.

12. References

- ACOPRA (2002). Standards for accreditation of physiotherapy programs at the level of higher education awards. Canberra: ACOPRA.
- Adams, C. L., Glavin, K., Hutchins, K., Lee, T., & Zimmerman, C. (2008). An evaluation of the internal reliability, construct validity, and predictive validity of the Physical Therapist Clinical Performance Instrument (PT CPI). *Journal of Physical Therapy Education*, 22(2), 42-50.
- Ajjawi, R., Loftus, S., Schmidt, H. G., & Mamede, S. (2009). Clinical reasoning: the nuts and bolts of clinical education. In C. Delany & E. Molloy (Eds.), *Clinical education in the health professions* (pp. 109-127). Sydney: Elsevier.
- Alexander, H. A. (1996). Physiotherapy student clinical education: The influence of subjective judgements on observational. *Assessment & Evaluation in Higher Education*, 21(4), 357.
- Allison, H., & Turpin, M. J. (2004). Development of the student placement evaluation form: a tool for assessing student fieldwork performance. *Australian Occupational Therapy Journal*, 51(3), 125-132.
- American Educational Research Association, American Psychological Association, National Council for Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, D.C: American Educational Research Association.
- Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Andrich, D. (1988). *Rasch models for measurement*. Newbury Park (CA): Sage.
- Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42 (Suppl 1), I7-I16.
- Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2003). RUMM2020 (Version 2020). Perth: RUMM laboratory.
- Australian Council for Safety and Quality in Health Care (2005). *National patient safety education framework*.
- Australian Council on Healthcare Standards (2002). *EQuIP Standards*.
- Australian Physiotherapy Council (2006). *Australian standards for physiotherapy*. Canberra: APC.

- Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation, 32*, 153-170.
- Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational Research Review, 2*, 114-129.
- Barbour, R. S. (2005). Making sense of focus groups. *Medical Education, 39*(7), 742-750.
- Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of General Internal Medicine, 20*(12), 1159-1164.
- Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1*, 307-310.
- Bogdan, R. B., & Biklin, S. K. (1998). *Qualitative research for education: An introduction to theory and methods* (3rd ed.). Needham Heights, MA.: Allyn and Bacon.
- Bond, T. G., & Fox, M. T. (2007). *Applying the Rasch Model. Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- Bosi Ferraz, M., Quaresma, M. R., Aquino, L. R. L., Atra, E., Tugwell, P., & Goldsmith, C. H. (1990). Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. *Journal of Rheumatology, 17*, 1022-1024.
- Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education, 31*(4), 399-413.
- Bourner, T., Martin, V., & Race, P. (1994). *Workshops that work: One hundred ideas to make your training events more effective*. London: McGraw-Hill Companies.
- Boyatzis, R. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks, CA: Sage.
- Bursari, J. O., Scherpbier, A. J., van der Vleuten, C. P. M., & Essed, G. G. (2006). A 2-day teacher-training programme for medical residents: investigating the impact on teaching performance. *Advances in Health Sciences Education, 11*, 133-144.
- Carliner, S. (2003). *Training design basics*. Baltimore: American Society for Training and Development.
- Carr, L. (2005). *Clinical education in physiotherapy*. Brisbane: Queensland Health.

- Chapman, J. (Ed.). (1998). *Developing professional judgement in health care: learning through the critical appreciation of practice*. Oxford: Butterworth-Heinemann.
- Checkland, P., & Holwell, S. (1998). Action research: Its nature and validity. *Systemic Practice and Action Research*, 11(1), 9-21.
- Christian, L. M., Parsons, N., & Dillman, D. A. (2009). Designing scalar questions for web surveys. *Sociological Methods and Research* 37(3), 393-425.
- Chudowsky, N., & Behuniak, P. (1998). Using focus groups to examine the consequential aspect of validity. *Educational measurement: Issues and Practice*, 17(4), 28-38.
- Cicchetti, D. V., Shoinralter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, 9(1), 31-36.
- Cliff, N., & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences*. Mahwah, New Jersey: Erlbaum.
- Coghlan, D., & Brannick, T. (2001). *Doing action research in your own organisation*. Thousand Oaks, CA: Sage.
- Cohen, J. W. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ.: Lawrence Erlbaum Associates.
- Conaghan, P. G., Emerton, M., & Tennant, A. (2007). Internal construct validity of the Oxford Knee Scale: evidence from Rasch measurement. *Arthritis Rheum*, 57(8), 1363-1367.
- Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *American Journal of Medicine*, 19(2), 166.e167-166.e116.
- Coote, S., Alpine, L., Cassidy, C., Loughnane, M., McMahon, S., Meldrum, D., et al. (2007). The development and evaluation of a Common Assessment Form for physiotherapy practice education in Ireland. *Physiotherapy Ireland*, 28(2), 6-10.
- Cote-Arsenault, D., & Morrison-Breedy, D. (1999). Practical advice for planning and conducting focus groups. *Nursing Research*, 48(5), 280-283.
- Cox, J. (1996). *Your opinion please!: How to build the best questionnaires in the field of education*. Thousand Oaks, California: Corwin Press.
- Cox, P. D., La, D. H., & Pappachan, M. (1999). Acquisition of physiotherapy competencies during clinical placements. *Physiotherapy Canada*, 51(2), 113-119.

- Crabtree, B., & Miller, W. (1999). A template approach to text analysis: Developing and using codebooks. In B. Crabtree & W. Miller (Eds.), *Doing qualitative research* (pp. 163-177). Newbury Park, CA: Sage.
- Cronbach, L. J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.
- Cross, V. (2001). Approaching consensus in clinical competence assessment: third round of a Delphi study of academics' and clinicians' perceptions of physiotherapy undergraduates. *Physiotherapy*, *87*(7), 341-350.
- Cross, V., & Hicks, C. (1997). What do clinical educators look for in physiotherapy students? *Physiotherapy*, *83*(5), 249-260.
- Cross, V., Hicks, C., & Barwell, F. (2001). Comparing the importance of clinical competence criteria across specialties: the impact on undergraduate assessment. *Physiotherapy*, *87*(7), 351-367.
- Daelmans, H. E., van der Hem-Stokroos, H. H., Hoogenboom, R. J., Scherpbier, A. J., Stehouwer, C. D., & van der Vleuten, C. P. (2005). Global clinical performance rating, reliability and validity in an undergraduate clerkship. *Netherlands Journal of Medicine*, *63*(7), 279-284.
- Davidson, M., & Keating, J. L. (2002). A comparison of five low back disability questionnaires: Reliability and responsiveness. *Physical Therapy*, *82*(1), 8.
- de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L., & Bouter, L. M. (2006). Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes* *4*(54).
- Delany, C., & Molloy, E. (2009). *Clinical Education in the Health Professions*. Sydney: Elsevier.
- Dickinson, T. L., & Zellinger, P. M. (1980). A comparison of the behaviourally anchored rating mixed standard scale formats. *Journal of Applied Psychology*, *65*, 147-154.
- Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement*, *44*, 61-66.
- Domingues, R. C. L., Amaral, E., & Zeferino, A. M. B. (2009). Global overall rating for assessing clinical competence: What does it really show? *Medical Education*, *43*, 883-886.

- Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education, 37*, 830-837.
- Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38*, 1006-1012.
- Drennan, J. (2003). Cognitive interviewing: Verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing, 42*(1), 57-63.
- Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*(4), 387-396.
- Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *Journal of American Medical Association, 287*(2), 226-235.
- Farbrigar, L. R., Krosnick, J. A., & MacDougall, B. L. (2005). Attitude measurement: Techniques for measuring the unobservable. In T. C. Brock & M. C. Green (Eds.), *Persuasion: Psychological insights and perspectives* (2nd ed., pp. 17-40). Thousand Oaks, California: Sage.
- Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. *Personnel Psychology, 35*, 105-117.
- Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 1-11.
- Fitzgerald, L. M., Delitto, A., & Irrgang, J. J. (2007). Validation of the Clinical Internship Evaluation Tool. *Physical Therapy, 87*(7), 844-860.
- Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment, 2*(14), 1-74.
- Friedman Ben-David, M. (2005). Principles of assessment. In J. A. Dent & R. M. Harden (Eds.), *A practical guide for medical teachers* (pp. 282-292). Edinburgh: Churchill Livingstone.
- Friedman, M., & Mennin, S. (1991). Rethinking critical issues in performance assessment. *Academic Medicine, 66*(7), 390-395.
- Goodwin, L. D. (2002). Changing Conceptions of Measurement Validity: An Update on the New Standards. *Journal of Nursing Education, 41*(3), 100-106.

- Govaerts, M. J., van der Vleuten, C. P., & Schuwirth, L. W. (2002). Optimising the reproducibility of a performance-based assessment test in midwifery education. *Adv Health Sci Educ Theory Pract*, 7(2), 133-145.
- Gravetter, F., & Wallnau, L. (2005). *Essentials of statistics for the behavioural sciences*. Pacific Grove: Wadsworth.
- Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, 11, 385-398.
- Haig, A., & Dozier, M. (2003b). BEME Guide No. 3: Systematic searching for evidence in medical education - Part 2: Constructing searches. *Medical Teacher*, 25(5), 463-484.
- Hambleton, R. K., & Jones, R. W. (1993). *Comparison of classical test theory and item response theory and their applications to test development*: University of Massachusetts.
- Hammond, J. A. (2009). Assessment of clinical components of physiotherapy undergraduate education: are there any issues with gender? *Physiotherapy*, 95, 266-272.
- Harden, R. M., Grant, G., Buckley, G., & Hart, I. R. (1999). BEME Guide No. 1: Best evidence medical education. *Medical Teacher*, 21(6), 553-562.
- Herbers, J. E. J., Noel, G. L., Cooper, G. S., Harvey, J., Pangaro, L. N., & Weaver, M. J. (1989). How accurate are faculty evaluations of clinical competence? *Journal of General Internal Medicine*, 4, 202-208.
- Higgs, J., & Bithell, C. (2001). Professional expertise. In J. Higgs & T. Titchen (Eds.), *Practice knowledge and expertise in the health professions*. (pp. 59-68). Oxford: Butterworth-Heinemann.
- Hobart, J., & Cano, S. (2009). Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment*, 13(12).
- Hobart, J. C., Cano, S. J., Zajicek, J. P., & Thompson, A. J. (2007). Rating scales as outcome measures for clinical trials in neurology: problems, solutions and recommendations. *Lancet Neurology*, 6, 1094-1105.
- Holden, R. R., Fekken, G. C., & Jackson, D. N. (1985). Structured personality test item characteristics and validity. *Journal of Research in Personality*, 19, 386-394.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30

- Hrachovy, J., Clopton, N., Baggett, K., Garber, T., Cantwell, J., & Schreiber, J. (2000). Use of The Blue MACS: acceptance by clinical instructor and self-reports of adherence. *Physical Therapy, 80*(7), 652-661.
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277-1288.
- Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75*, 334-349.
- Jensen, M. P., Turner, J. A., & Romano, J. M. (1994). What is the maximum number of levels needed in pain intensity measurement? *Pain, 58*, 387-392.
- Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527-535.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17.
- Kassirer, J., Wong, J., & Kopelman, R. (2010). *Learning clinical reasoning*. Philadelphia, PA: Lippincott Williams and Wilkins.
- Keating, J. L., & Matyas, T. (1998). Unreliable inferences from reliable measurements. *Australian Journal of Physiotherapy, 44*(1), 5-10.
- Keen, A. J., Klein, S., & Alexander, D. A. (2003). Assessing the communication skills of doctors in training: reliability and sources of error. *Advances in Health Sciences Education, 8*, 5-16.
- Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing Behaviourally Anchored Rating Scales (BARS) and other rating formats. *Personnel Psychology, 34*(2), 263-289.
- Kitzinger, J. (1995). Qualitative research: Introducing focus groups. *British Medical Journal, 311*, 299-302.
- Kivunja, C. (2009). *Celebrating innovative qualitative data analysis using Leximancer*. Paper presented at the 4th International Postgraduate Research Conference.
- Kline, P. (1986). *A handbook of test construction*. London: Methuen.
- Knight, A., Carrese, J., & Wright, S. (2007). Qualitative assessment of the longterm impact of a faculty development programme in teaching skills. *Medical Education, 41*(6), 592-600.

- Knowles, M. S., Holton, E. F., & Swanson, R. A. (2005). *The adult learner: The definitive classic in adult education and human resource development* (6th ed.). Burlington, MA.: Elsevier.
- Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees. *Journal of American Medical Association, 302*(12), 1316-1326.
- Kolb, D. A. (1984). *Experiential learning: Experience as a source of learning and development*. Englewood Cliffs, NJ.: Prentice Hall.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measurement surveys. *Applied Cognitive Psychology, 5*, 213-236.
- Krosnick, J. A., & Berent, M. K. (1993). Comparisons of party identification and policy preferences: The impact of survey question format. *American Journal of Political Science, 37*, 941-964.
- Krosnick, J. A., & Farbrigar, L. R. (1997). *Designing good questionnaires effectively*. New York: Oxford University Press.
- Krosnick, J. A., & Presser, S. (2009). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed.). San Diego, California: Elsevier.
- Krueger, R., & Casey, M. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). California: Sage.
- Lai, J.-S., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation and the Health Professions, 28*(3), 283-294.
- Lam, T. C. M., & Kolic, M. (2008). Effects of semantic incompatibility on rating response. *Applied Psychological Measurement, 32*(3), 248-260.
- Lamoureux, E. L., Pallant, J. F., Pesudovs, K., & Hassell, J. B. (2006). The impact of vision impairment questionnaire: an evaluation of its measurement properties using Rasch analysis. *Investigative Ophthalmology and Visual Science, 47*(11), 4732-4741.
- Landis, J. R., & Koch, C. G. (1977). A one-way components of variance model for categorical data. *Biometrics, 33*, 671-679.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

- Likert, R. A. (1952). A technique for the development of attitude scales. *Educational and Psychological Measurement*, 12, 313-315.
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7, 328.
- Logemann, H. L. (2006). *An item response theory analysis of the Physical Therapist Manual for the Assessment of Clinical Skills (PT MACS) and the physical therapist Clinical Performance Instrument (CPI)*. Unpublished Ph.D., Texas Woman's University, United States -- Texas.
- Loomis, J. (1985a). Evaluating clinical competence of physical therapy students: assessing the reliability, validity and usability of a new instrument... part 2. *Physiotherapy Canada*, 37(2), 91-98.
- Loomis, J. (1985b). Evaluating clinical competence of physical therapy students: the development of an instrument... part 1. *Physiotherapy Canada*, 37(2), 83-89.
- MacLellan, E. (2001). Assessment for learning: Different perceptions of tutors and students. *Assessment and Evaluation in Higher Education*, 26(4), 307-318.
- Malhotra, N., Krosnick, J. A., & Thomas, R. K. (2007). Optimal design of branching questions to measure bipolar constructs. Unpublished report. Stanford University.
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Masters, G. N., & Keeves, J. P. (1999). *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon.
- Masters, G. N., & Wilson, M. (1997). *Developmental assessment*. Berkeley, California: University of California.
- Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, 11, 49-53.
- Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, 11, 49-53.
- McAllister, S. (2005). *Competency based assessment of speech pathology students' performance in the workplace*. The University of Sydney, Sydney.

- McAllister, S., Lincoln, M., Ferguson, A., & McAllister, L. (2010). Issues in developing valid assessments of speech pathology students' performance in the workplace. *International Journal of Language Communication Disorders, 45*(1), 1-14.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.
- McIlroy, J. H., Hodges, B., McNaughton, N., & Regehr, G. (2002). The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. *Academic Medicine, 77*(7), 725-728.
- McKelvie, S. J. (1978). Graphic rating scales - how many categories? *British Journal of Psychology, 69*, 185-202.
- McLeod, P. J., Meagher, T. W., Steinert, Y., & Boudreau, D. (2000). Using focus groups to design a valid questionnaire. *Academic Medicine, 75*, 671.
- Meldrum, D., Lydon, A., Loughnane, M., Geary, F., Shanley, L., Sayers, K., et al. (2008). Assessment of undergraduate physiotherapist clinical performance: Investigation of educator inter-rater reliability. *Physiotherapy, 94*, 212-219.
- Melia, K. M. (1997). Producing 'plausible stories': Interviewing student nurses. In G. Miller & R. Dingwall (Eds.), *Context and method in qualitative research* (pp. 26-36). London: Sage.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.
- Messick, S. (1994). The interplay of evidence and consequence in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.
- Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*, 5-8.
- Messick, S. (1995b). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.
- Messick, S. (1996). Validity of Performance Assessment. In G. Phillips (Ed.), *Technical Issues in Large-Scale Performance Assessment*. Washington DC: National Center for Educational Statistics.

- Michell, J. (2008). Is psychometrics pathological science? *Measurement - Interdisciplinary Research and Perspectives*, 6, 7-24.
- Miller, G. A. (1956). The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Bulletin*, 63, 81-97.
- Miller, G. E. (1990). The assessment of clinical skills/competence/performance *Academic Medicine*, 65, 563-567.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann Intern Med*, 151, 264-269.
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., et al. (2010). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual of Life Research*, February.
- Molloy, E. (2006). *Insights into the formal feedback culture in physiotherapy clinical education*. Monash, Melbourne.
- Morgan, D. L. (1988). *Focus groups as qualitative research*. London: Sage.
- Morgan, P. J., Cleave-Hogg, D., & Guest, C. B. (2001). A comparison of global ratings and checklist scores from an undergraduate assessment using an anesthesia simulator. *Academic Medicine*, 76(10), 1053-1055.
- Morris, J. (2006). Audit of use of a common undergraduate physiotherapy clinical assessment form. *International Journal of Therapy & Rehabilitation*, 13(9), 407-413.
- Morton, J., Cumming, A., & Cameron, H. (2007). Performance-based assessment in undergraduate medical education. *Clinical Teacher*, 4, 36-41.
- Munshi, J. (1990). *A method for constructing Likert scales: Research report*. California: Sonoma State University.
- Newble, D. (2004). Techniques for measuring clinical competence: Objective structured clinical examinations. *Medical Education*, 38, 199-203.
- Newble, E. D., Jolly, B., & Wakeford, R. (1994). The certification and recertification of doctors: Issues in the assessment of clinical competence. *BMJ*, 309(1096).
- Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education*, 37(5), 464-469.

- Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: a method for assessing clinical skills. *Annals of Internal Medicine, 138*(6), 476-481.
- Norman, G. R., Sloan, J. A., & Wywich, K. W. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care, 41*, 582-592.
- Notzer, N., & Abramovitz, R. (2008). Can brief workshops improve clinical instruction? *Medical Education, 42*, 152-156.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*, 1-18.
- Pallant, J. (2005). *SPSS survival manual* (2nd ed.). Sydney: Allen and Unwin.
- Pallant, J. F. (2007). *Step by step guide to doing Rasch analysis using RUMM2020. Training materials*. Melbourne: Swinburne University.
- Pallant, J. F. (2010). Targeting of items in Rasch analysis. In M. Dalton (Ed.).
- Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(Pt 1), 1-18.
- Papadakis, M., Hodgson, C., Teherani, A., & Kohatsu, N. (2004). Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. *Academic Medicine, 79*, 244-249.
- Pelgrim, E. A. M., Kramer, A. W. M., Mokkink, H. G. A., van den Elsen, L., Grol, R. P., & van der Vleuten, C. P. M. (2010). In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in Health Sciences Education, Education,*
- Portney, L., & Watkins, M. (1993). *Foundations of clinical research : applications to practice*. Norwalk, Conn.: Appleton and Lange.
- Prescott-Clements, L., van der Vleuten, C. P., Schuwirth, L. W., Hurst, Y., & Rennie, J. S. (2008). Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). *Medical Education, 42*(5), 488-495.
- Preston, C. C., & Coleman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power and respondent preferences. *Acta Psychologica, 104*, 1-15.

- Race, P. (2002). *Two thousand tips for trainers and staff developers*. London: Taylor and Francis Inc.
- Rankin, G., & Stokes, M. (1998). Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical Rehabilitation*, *12*, 187-199.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institute.
- Rethans, J. J., Norcini, J. J., Baron-Maldonado, M., Blackmore, D., Jolly, B. C., LaDuca, T., et al. (2002). The relationship between competence and performance: implications for assessing practice performance. *Medical Education*, *36*(10), 901-909.
- Rheault, W., & Coulson, E. (1991). Use of the Rasch model in the development of a clinical competence scale. *Journal of Physical Therapy Education*, *5*(1), 10-13.
- Ries, J. D., Echternach, J. L., Nof, L., & Gagnon Blodgett, M. (2009). Test-retest reliability and minimal detectable change scores for the timed "up and go" test, the six minute walk test and gait speed in people with alzheimer disease. *Physical Therapy*, *89*(6), 569-579.
- Ringsted, C., Ostergaard, D., Ravn, L., Pedersen, J. A., Berlac, P. A., & Van Der Vleuten, C. P. M. (2003). A feasibility study comparing checklists and global rating forms to assess resident performance in clinical skills. *Medical Teacher*, *25*(6), 654-658.
- Rogoff, B. (1990). *Cognitive Development in Social Context*. New York: Oxford University Press.
- Rousson, V., Gasser, T., & Seifert, B. (2002). Assessing intrarater, interrater and test-retest reliability of continuous measurements. *Statistics in Medicine*, *21*, 3431-3446.
- Rowland-Morrin, P. A., Burchard, K. W., Garb, J. L., & Coe, N. P. (1991). Influence of effective communication by surgery students on their oral examination scores. *Academic Medicine*, *66*, 169-171.
- Sandelowski, M. (2009). What's in a name? Qualitative description revisited. *Research in Nursing and Health*, *33*(1), 77-84.
- Schuman, H., & Presser, S. (1981). *Questions and answers*. New York: Academic Press.
- Schwarz, N., Knauper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*, 570-582.

- Seymour, R. A., Simpson, J. M., Charlton, J. E., & Phillips, M. E. (1985). An evaluation of length and end-phrase of visual analog scales in dental pain. *Pain, 21*, 177-185.
- Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27*(2), 4-13.
- Shavelson, R. J., Gao, X., & Baxter, G. (1993). *Sampling variability in performance assessments*. Santa Barbara: National Center for Research on Evaluation, Standards and Student Testing, University of California.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.
- Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.
- Smith, E. V. (2002). Detecting and evaluation of the impact of multidimensionality using item fit statistics and principal components analysis of residuals. *Journal of Applied Measurement, 3*, 205-231.
- Southgate, L., Hays, R. B., Norcini, J., Mulholland, H., Ayers, B., Woolliscroft, J., et al. (2001). Setting performance standards for medical practice: a theoretical framework. *Medical Education, 35*(5), 474-481.
- Spector, P. E. (1992). Summated rating scale construction: An introduction. In M. S. Lewis-Beck (Ed.), *Quantitative applications in the social sciences*. Newbury Park, CA: Sage Publications Inc.
- Stickley, L. A. (2002). *Validity and reliability of a clinical education performance tool for student physical therapists*. Unpublished Ph.D., Texas Tech University.
- Stickley, L. A. (2005). Content validity of a clinical education performance tool: the Physical Therapist Manual for the Assessment of Clinical Skills. *Journal of Allied Health, 34*(1), 24-30.
- Stratford, P. W. (2004). Getting more from the literature: estimating the standard error of measurement from reliability studies. *Physiotherapy Canada, 56*, 27-30.
- Stratford, P. W., Binkley, F. M., Solomon, P., Finch, E., Gill, C., & Moreland, J. (1996). Defining the minimum level of detectable change for the Roland-Morris Questionnaire. *Physical Therapy, 76*(4), 359-365.

- Straube, D., & Campbell, S. K. (2003). Rater discrimination using the visual analog scale of the Physical Therapist Clinical Performance Instrument. *Journal of Physical Therapy Education, 17*(1), 33-38.
- Streiner, D. L., & Norman, G. R. (2003). *Health Measurement Scales. A practical guide to their development and use.* (3rd ed.). New York: Oxford University Press.
- Task Force for the Development of Student Clinical Performance Instruments (2002). The development and testing of APTA clinical performance instruments. *Physical Therapy, 82*(4), 329-353.
- Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism, 57*(8), 1358-1362.
- Tennant, A., McKenna, S. P., & Hagell, P. (2004). Applications of Rasch analysis in the development and application of quality of life instruments. *Value Health, 7*(Suppl 1), S22-66.
- Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters! (a tale of two Smiths). *Rasch Measurement Transactions, 20*(1), 1048-1051.
- Teresi, J. (2001). Statistical methods for examination of Differential Item Functioning (DIF) with applications to cross-cultural measurement of functional, physical and mental health. *Journal of Mental Health and Aging, 7*, 31-40.
- Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology, 60*(1), 34-42.
- Tourangeau, R., Rips, L., & Rasinski, K. A. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.
- Tsuda, H., Low, S., & Vlad, G. (2007). A description of comments written by clinical instructors on the Clinical Performance Instrument. *Journal of Physical Therapy Education, 21*(1), 56-62.
- Van der Vleuten, C. (1996a). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education Theory and Practice, 1*(1), 41-67.
- van der Vleuten, C. (1996b). Making the best of the "long case". *Lancet, 347*(9003), 704-705.

- van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *British Medical Journal*, 321(7270), 1217-1219.
- Vygotsky, L. (1986). *Thought and Language*. Cambridge: Massachusetts Institute of Technology Press.
- Wang-Cheng, R. M., Fulkerson, P. K., Barnas, G. P., & Lawrence, S. L. (1995). Effect of student and preceptor gender on clinical grades in an ambulatory care clerkship. *Academic Medicine*, 70, 324-326.
- Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001a). Assessment of clinical competence. *Lancet*, 357(9260), 945-949.
- Wass, V., van der Vleuten, C., Shatzer, J., & Jones, R. (2001b). Assessment of clinical competence. *The Lancet*, 357, 945-367.
- Watkins, M. (2000). Monte Carlo PCA for Parallel Analysis [Computer software]. . Phoenix, AZ: Ed & Psych Associates.
- Watson, D. (1988). The vicissitudes of mood measurement: Effects of varying descriptors, time frames and response formats on measures of positive and negative affect. *Journal of Personality and Social Psychology*, 55, 341-356.
- Waugh, R. (2005). An analysis of dimensionality using factor analysis (true score theory) and Rasch measurement: what is the difference? Which method is better? *Journal of Applied Measurement*, 6, 80-99.
- Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, 19(1), 231-240.
- Willis, G. B. (2005). *Cognitive interviews: A tool for improving questionnaire design*. Thousand Oaks, California: Sage.
- Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: part II - validation activities. *Journal of Applied Measurement*, 8(2), 204-234.
- Wolfe, E. W., & Smith, E. V., Jr (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I - instrument development tools. . *Journal of Applied Measurement*, 8(1), 97-123.

- Wong, D. L., & Baker, C. (1988). Pain in children: comparison of assessment scales. *Pediatric Nursing, 14*, 9-17.
- Wright, B. D. (1996a). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*(1), 3-24.
- Wright, B. D. (1996b). Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*, 509-511.
- Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.
- Wright, B. D., & Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement, 1*, 83-106.
- Yates, J., & James, D. (2006). Predicting the "strugglers": a case-control study of students at Nottingham University Medical School. *British Medical Journal, 332*, 1009-1013.

13. Appendices (refer to volume 2)