The Assessment of Physiotherapy Practice (APP) is a valid measure of professional competence of physiotherapy students: a cross-sectional study with Rasch analysis

Megan Dalton[1,3], Megan Davidson[2] and Jenny Keating[3]

[1]Griffith University, [2]School of Allied Health, La Trobe University, [3]School of Primary Health Care, Monash University
Australia

Question: Is the Assessment of Physiotherapy Practice (APP) a valid instrument for the assessment of entry-level competence in physiotherapy students? **Design**: Cross-sectional study with Rasch analysis of initial (n = 326) and validation samples (n = 318). Students were assessed on completion of 4, 5, or 6-week clinical placements across one university semester. **Participants**: 298 clinical educators and 456 physiotherapy students at nine universities in Australia and New Zealand provided 644 completed APP instruments. **Results**: APP data in both samples showed overall fit to a Rasch model of expected item functioning for interval scale measurement. Item 6 (Written communication) exhibited misfit in both samples, but was retained as an important element of competence. The hierarchy of item difficulty was the same in both samples with items related to professional behaviour and communication the easiest to achieve and items related to clinical reasoning the most difficult. Item difficulty was well targeted to person ability. No Differential Item Functioning was identified, indicating that the scale performed in a comparable way regardless of the student's age, gender or amount of prior clinical experience, and the educator's age, gender, or experience as an educator, or the type of facility, university, or clinical area. The instrument demonstrated unidimensionality confirming the appropriateness of summing the scale scores on each item to provide an overall score of clinical competence and was able to discriminate four levels of professional competence (Person Separation Index = 0.96). Person ability and raw APP scores had a linear relationship (r^2 = 0.99). **Conclusion**: Rasch analysis supports the interpretation that a student's APP score is an indication of their underlying level of professional competence in workplace practice. **[Dalton M, Davidson M, Keating J (2011) The Assessment of Physiotherapy Practice (APP) is a valid measure of professional competence of physiotherapy students: a cross-sectional study with Rasch analysis. *Journal of Physiotherapy* 57: 239–246]**

Key words: Educational measurement, Professional competence, Clinical competence, Physical therapy (specialty)

Introduction

Workplace-based learning and assessment is an essential component of physiotherapy and other health professional education programs. Professional competence includes understanding and dealing with highly variable circumstances and assessment is therefore difficult to standardise across students (Rethans et al 2002). Controlled assessments such as Objective Structured Clinical Examinations and the use of standardised patients have been developed in response to concerns regarding standardised and reliable measurement of student competencies. While assessment reliability may be enhanced by standardised 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).

A solution to this complexity is to monitor students over a sufficient 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 assessment of professional competence of physiotherapy students in the workplace is the assessment approach used within all Australian and New Zealand physiotherapy programs. Clinical educators (registered physiotherapists) generally rate a student's performance on a set of items on completion of a 4, 5, or 6-week block of supervised workplace practice. If valid interpretations of such scores are to be made, the assessment instrument must be both psychometrically sound and educationally informative (Prescott-Clements et al 2008, Streiner and Norman 2003). These requirements were fundamental considerations in the development and evaluation of the Assessment of Physiotherapy Practice (APP) instrument (Dalton et al 2009), which has been adopted in all but one Australian and all New Zealand entry-level programs.

The development of the APP was guided by the framework of Wilson (2005). An initial item pool was constructed from all available assessment instruments and reduced by removing redundancy and applying criteria related to good

What is already known on this topic: Assessment of clinical competence under controlled conditions of practical examinations may not be an adequate surrogate for performance in clinical practice. A standard assessment tool is needed for physiotherapy students on clinical placements.

What this study adds: The Assessment of Physiotherapy Practice (APP) is a valid measure of professional competence of physiotherapy students. It is appropriate to sum the scale scores on each item to provide an overall score of clinical competence. The APP performs in a comparable way regardless of the characteristics of the student, the clinical educator, or the clinical placement.

item design. The test content development included input and collaboration from physiotherapy educators across Australia and New Zealand. The iterative cycles included a pilot trial and two field test stages. A detailed description of these stages is presented in Figure 1 (see eAddenda for Figure 1). Continuous refinement of the instrument based on qualitative and quantitative evaluation occurred throughout each stage (Coghlan and Brannick 2001). There were three phases of workplace-based testing – a pilot trial and two field tests (Dalton et al 2009). This paper reports the results of the second field test.

Rasch analysis of data was used at each stage of testing the APP. This statistical model calibrates the difficulty of items and the ability of persons on a common scale with interval-level units called logits (log-odds units) (Bond and Fox 2007, Rasch 1960). Rasch analysis provides validity evidence based on instrument internal structure. It enables analysis of unidimensionality (considered an essential quality of an additive scale) and the targeting of item difficulty to the persons' abilities (Bond and Fox 2007). Rasch analysis also enables assessment of the functioning of the rating scale when applied to students with different characteristics (eg, age and gender) or applied by assessors with different characteristics (eg, years of experience as a clinical educator). If data fit a Rasch model, a number of qualities should be evident in the data. Items should present a stable hierarchy of difficulty. It should be easy to achieve high scores on easy items and difficult on hard items, with items in between ranking in a predictable way. An instrument with these properties would make the user confident that a student who achieved a higher total score was able to cope with the more difficult, as well as the easier, challenges. Educators could identify challenging items and appropriate educational support could be developed to help students achieve these more challenging aspects of practice. Further detail on the methods of Rasch analysis and the applicability of its results in the clinical environment is provided in an excellent paper by Tennant and Conaghan (2007).

The aim of this study was to ascertain whether the APP instrument is a valid measure of professional competence of physiotherapy students when tested using the Rasch measurement model. Therefore the specific research questions were:

1. Is the APP a unidimensional measure of the professional competence of physiotherapy students?
2. What is the hierarchy of difficulty of items from easiest to hardest?
3. Is there any evidence of differential item functioning, which indicates the scale exhibits item bias?
4. Are the APP items appropriately targeted for the student population?

Method

This was a cross-sectional study using Rasch analysis of two samples (n = 326 and n = 318). Students were assessed at completion of clinical placements across one university semester in 2008. Approval was obtained from the human ethics committee of each participating university.

The APP (Version 4) used in this final field trial comprised 20 items, presented in Appendix 1 (see the eAddenda for Appendix 1). Each of the 20 items has the response options 0 = *infrequently/rarely demonstrates performance indicators,* 1 = *demonstrates few performance indicators to an adequate standard,* 2 = *demonstrates most performance indicators to an adequate standard,* 3 = *demonstrates most performance indicators to a good standard,* 4 = *demonstrates most performance indicators to an excellent standard,* and *not assessed.* A rating of 0 or 1 indicates that a minimum acceptable standard has not been achieved for that item. A global rating scale of overall performance (not adequate, adequate, good, excellent) is also completed by the educator, but this item does not contribute to the APP score. Examples of performance indicators for each item are provided on the reverse of the APP. A total raw score for the APP ranges from 0 to 80, and can be transformed to a 0 to 100 scale by dividing the raw score by the total number of items scored (ie, excluding any items that were not assessed) and multiplying the result by 100.

Participants

Students enrolled in entry-level physiotherapy programs from 9 universities in Australia and New Zealand were assessed by educators using the APP on completion of a 4, 5, or 6-week full-time clinical placement block scheduled across one university semester. The placements occurred during the last 18 months of the students' physiotherapy program and represented diverse areas of physiotherapy practice including musculoskeletal, cardiorespiratory, neurological, paediatric, and gerontological physiotherapy.

Recruitment procedures optimised representation of physiotherapy clinical educators by location (metropolitan, regional/rural, and remote), clinical area of practice, years of experience as a clinical educator, and organisation (private, public, hospital based, community based, and non-government).

Field test procedure

Prior to commencement of clinical placements, educators and students were sent an information sheet and consent form and invited to participate. Data were excluded from analysis if either the student or their clinical educator did not consent to participate in the research. All clinical educators received training in the use of the APP through attendance at a 4-hour workshop, access to the APP resource manual, or both. Compulsory workshop attendance for all clinical educators participating in the field test was not feasible in the authentic clinical education environment where face-to-face training opportunities are constrained by geographical, workload, and financial considerations. During the trial a member of the research group was available to answer questions by phone or email. Students were educated in the assessment process and use of the APP instrument using a standardised presentation prior to placements commencing and information about the APP was included in each university's student clinical education manual.

Data management and analysis

On completion of each placement the completed APP forms were returned by mail, de-identified, and entered into a spreadsheet. Data were analysed with RUMM2020 software using a partial credit model (Andrich et al 2003). The analysis tested the overall fit of data to the model, the overall and individual item and person fit, item threshold order, targeting, item difficulty, person separation, differential item functioning, and dimensionality.

The Rasch measurement model

Conversion of ordinal data to interval level measurement data: The current approach in workplace-based assessment is to score a physiotherapy student's performance on a rating scale across items that sample behaviours considered essential for professional competence. 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 and Keats 2003, Streiner and Norman 2003). Rasch modeling enables the abstraction of equal units of measurement from raw (ordinal data) scores on items of an assessment tool. These can be calibrated and then used with confidence to measure and quantify attributes such as competence in physiotherapy practice (Bond and Fox 2007). This conversion facilitates appropriate interpretation of differences between individuals and tallying of converted scores provides interpretable total scores.

Functioning of items: In this study the construct of interest was competence to practice physiotherapy. If scores for items fit a Rasch model, a number of qualities should be evident in the data. Items should present a stable hierarchy of difficulty. It should be easy to achieve high scores on some items and difficult on others, with items in-between ranking in a reliable way. An instrument with these properties would make the user confident that a student who achieved a higher total score was able to cope with the more difficult, as well as the easier, challenges. Educators could identify challenging items and appropriate educational support could be developed to help students achieve these more challenging targets.

Item bias: A scale that fits a Rasch model 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 be scored significantly differently (Lai et al 2005). Rasch analysis enables assessment of item bias through investigation of Differential Item Functioning. In the development of the APP, the research team was particularly interested to determine whether the scale performed in a comparable way regardless of the student's age, gender, or the total number of weeks of clinical experience, the educator's age, gender, or experience as an educator, the type of facility where the clinical placement occurred, the university that delivered the student's education, or the clinical area.

Dimensionality: One of the primary tenets underpinning Rasch analysis is the concept of unidimensionality. If the scale scores on each item of the APP are to be added together to provide a total score representing an overall level of professional competence, Rasch analysis should indicate a scale that is unidimensional, a scale that measures one construct. Unidimensionality was explored using the independent t-test procedure (Tennant and Pallant 2006).

Targeting of instrument: It is important, particularly in clinical practice, that the assessment items are appropriately targeted for the population being assessed. Poorly targeted measures result in floor or ceiling effects, and this would mean that either very weak or very strong students may not be graded appropriately. Rasch modeling provides an indication of the match between the item difficulty and the abilities of people in the sample. A well-targeted scale would have a mean person location around zero (Tennant and Conaghan 2007).

Functioning of the rating scale: Rasch analysis generates a person separation index that provides an indication of the internal consistency of the scale and the power of the instrument to discriminate amongst respondents with different levels of professional competence. A minimum person separation index of 0.70 and 0.85 is required for group and individual use respectively (Tennant and Conaghan 2007).

Rasch analysis also enables investigation of difficulty that clinical educators may have in discriminating between different levels on the 0–4 rating scale. For a good fit to the model it is expected that for any item, student with high levels of the attribute (professional competence indicated by total scores) would typically achieve a higher item score than individuals with low levels of the attribute. In Rasch analysis this is demonstrated by an ordered set of response thresholds for each item. Ordered thresholds indicate that the respondents (ie, clinical educators) use the response categories (ie, scoring scale) in a manner consistent with the level of the trait (ie, competence) being measured. This occurs when the educators consistently discriminate between response options in a predictable way.

Results

A total of 644 APP assessments from 456 students were returned by 298 clinical educators. Tables 1 and 2 present the characteristics of the participating students and educators. Table 3 presents the characteristics of the APP forms received. The mean APP total score was 61 (SD 12, range 16–80). If converted to the 0–100 scale, this equates to a mean total score of 76 (SD 15, range 20–100). All 5 points on the rating scale were used for the majority of items. Missing data was rare (0.4% of all data points) and 0.2% of all items were rated as *not assessed.*

Data were randomly divided into two samples. Sample 1 was used for model development (n = 326) and sample 2 for model validation (n = 318). The data were stratified before randomisation 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 type (undergraduate, graduate entry).

Overall model fit: The item-trait interaction chi-square statistic for Sample 1 was 65.1 (df = 80, p = 0.88) and 100 (df = 80, p = 0.57) for Sample 2. The chi-square probability values for Sample 1 (p = 0.88) and Sample 2 (p = 0.57) indicated adequate fit between the data and the model.

Overall item and person fit: The residual mean value for items for Sample 1 was –0.33 (SD 1.71), and for Sample 2 was –0.32 (SD 1.73), indicating some misfit of items. The residual mean value for persons for Sample 1 was –0.26 (SD 1.19) and for Sample 2 was –0.19 (SD 1.13), indicating no misfit of persons in either sample.

Individual item and person fit: In both samples, Item 6 (Demonstrates clear and accurate written documentation) exhibited a positive item fit residual above +2.5, suggesting

Table 1. Characteristics of participating students.

Characteristic	n = 456
Age *(yr)*, mean (SD)	23 (3)
Age *(yr)*, range	20–48
Gender, n female *(%)*	301 (66)

Table 2. Characteristics of participating clinical educators.

Characteristic	n = 298
Age *(yr)*, mean (SD)	34 (8)
Age *(yr)*, range	22–60
Gender, n female *(%)*	215 (72)
Experience as a clinical educator *(yr)*, mean (SD)	6 (5)
Experience as a clinical educator *(yr)*, range	0–34
Self-rated level of experience as a clinical educator, n (%)	
no experience	13 (4)
some experience	58 (19)
average experience	88 (30)
above average experience	86 (29)
very experienced	53 (18)

poor discrimination. None of the items exhibited a significant chi-square value (Table 4). To investigate if the misfit of Item 6 was contributing to the overall item misfit to the model, Item 6 was removed from each sample and Rasch analysis repeated. The residual mean value for overall item fit changed from –0.33 (SD 1.71) to –0.33 (SD 1.53) in Sample 1 and from –0.33 (SD 1.73) to –0.32 (SD 1.51) in Sample 2. The reduction in score variability indicated a small improvement in the overall fit of items to the model.

Threshold order: There were no disordered thresholds for any of the 20 items in either Sample 1 or 2. The threshold map for Sample 1 is illustrated in Figure 2.

Targeting: The average person location in both samples was close to zero (–0.06) indicating that overall the item difficulty was well targeted to the students' abilities. The person-item threshold graph (Figure 3) presents the distribution of the students (top half of the graph) and item thresholds (bottom half of the graph) on a logit scale for Sample 1. This graph shows 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. There appears to be an even spread of item thresholds across the full range of student abilities, suggesting effective targeting of APP items. Similar results were seen for the first field test. At the far right end of the X-axis, 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.

Hierarchy of item difficulty: The sequence or hierarchy of average difficulty of the 20 items on the APP for both samples is presented in Table 4. In both samples, items representing professional behaviour and communication were amongst the least difficult items whereas the most difficult items related to analysis and planning, progressing intervention, and applying evidence-based practice.

Table 3. Sources and characteristics of the APP forms received.

Characteristic	n = 644
Responder burden	
duration to complete *(min)*, mean (SD)	29 (19)
duration to complete *(min)*, range	8–120
Clinical area, (%)	
musculoskeletal	32
neurological	25
cardiorespiratory	23
paediatric	6
specialty[a]	5
unknown	8
Patient age group, (%)	
children (0 to 12 yr)	4
adolescents (13 to 20 yr)	3
adults (21 to 65 yr)	51
older persons (> 65 yr)	36
unknown	5
Type of facility[b], (%)	
public hospital	54
community based services	9
private hospital	7
non-government organisation	6
private practice	3
unknown	21
University program	
Monash University (Victoria, Australia)	28
Griffith University (Queensland, Australia)	20
La Trobe University (Victoria, Australia)	19
James Cook University (Queensland, Australia)	7
Curtin University (Western Australia, Australia)	7
The University of Sydney (New South Wales, Australia)	6
Charles Sturt University (New South Wales, Australia)	3
Otago University (New Zealand)	1
Auckland University of Technology (New Zealand)	1
unknown	7

[a]spinal injuries, burns, women's health, oncology, mental health, hand therapy, plastic surgery, [b]n = 423

Person separation index: The person separation index was 0.95 for Sample 1 and 0.96 for Sample 2, indicating that the APP is able to discriminate at least four levels of performance.

Differential item functioning: The presence of item bias was explored by analysis of differential item functioning with a Bonferroni-adjusted *p* value of 0.0025. No significant

Table 4. Individual item fit of 20 APP items to the Rasch model: sample 1 (n = 326) and sample 2 (n = 318), with items ordered from least to most difficult.

Sample 1 (n = 326) APP item	Location	Standard error	Fit residual	DF	Chi-square[a]	*p*	Sample 2 (n = 318) APP item	Location	Standard error	Fit residual	DF	Chi–square[a]	*p*
1	–2.088	0.136	0.796	306.73	1.94	0.746	1	–1.824	0.128	1.104	280.98	5.765	0.217
3	–1.296	0.121	2.267	306.73	3.723	0.444	3	–1.516	0.119	1.726	280.98	4.587	0.332
2	–0.997	0.137	1.418	306.73	6.152	0.188	2	–0.532	0.124	–0.887	280.05	1.597	0.809
6	–0.647	0.121	4.479	306.73	13.939	0.007	5	–0.486	0.129	1.219	280.98	1.105	0.893
7	–0.455	0.116	–1.078	306.73	1.161	0.884	6	–0.466	0.112	3.671	280.05	0.665	0.955
4	–0.174	0.121	–0.358	306.73	3.856	0.425	7	–0.451	0.117	0.478	280.98	2.165	0.705
5	–0.154	0.114	0.46	306.73	1.759	0.779	4	–0.133	0.11	–2.121	280.98	1.462	0.833
20	–0.073	0.119	–1.85	306.73	3.346	0.501	20	–0.106	0.111	–1.863	280.98	5.841	0.211
14	–0.025	0.122	–0.539	305.79	1.537	0.820	14	–0.094	0.123	–0.724	280.98	8.107	0.087
15	0.286	0.114	–0.235	306.73	3.295	0.509	15	–0.011	0.119	1.108	280.98	2.286	0.683
16	0.297	0.115	–1.105	306.73	1.052	0.901	9	0.01	0.111	–0.14	280.05	3.503	0.477
18	0.401	0.122	–1.308	306.73	4.864	0.301	16	0.062	0.112	1.266	278.17	5.059	0.281
8	0.440	0.112	–2.54	306.73	6.308	0.177	18	0.11	0.119	–2.612	280.98	1.094	0.895
9	0.496	0.114	–2.166	306.73	3.993	0.406	8	0.158	0.111	0.741	273.49	8.757	0.067
11	0.508	0.114	–4.023	305.79	6.733	0.150	13	0.32	0.11	–1.285	278.17	3.389	0.494
13	0.509	0.113	2.14	304.85	3.857	0.425	19	0.321	0.112	–2.317	280.98	11.03	0.026
19	0.514	0.113	–0.178	304.85	2.162	0.706	11	0.719	0.111	–3.286	279.11	6.669	0.154
12	0.716	0.116	0.165	306.73	1.365	0.850	17	0.784	0.112	–1.008	280.98	8.001	0.091
17	0.845	0.115	–1.455	305.79	2.27	0.686	10	0.847	0.111	–0.61	280.05	7.732	0.101
10	0.896	0.115	–2.096	306.73	7.796	0.099	12	1.016	0.115	–0.827	280.05	3.024	0.553

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. DF = degrees of freedom, [a]Chi-square degrees of freedom was 4 for all items

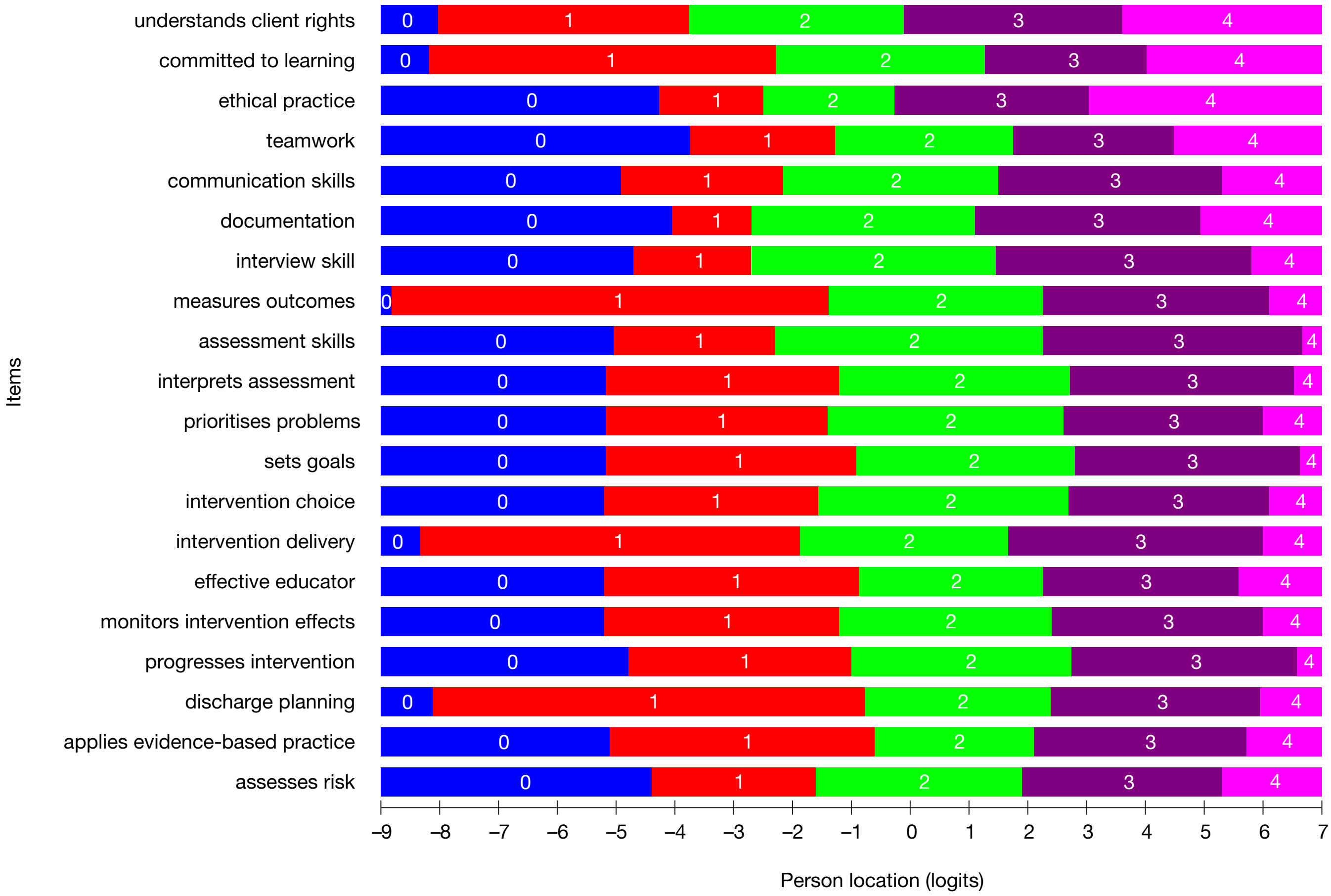

Figure 2. Threshold map in Sample 1 (n = 326).

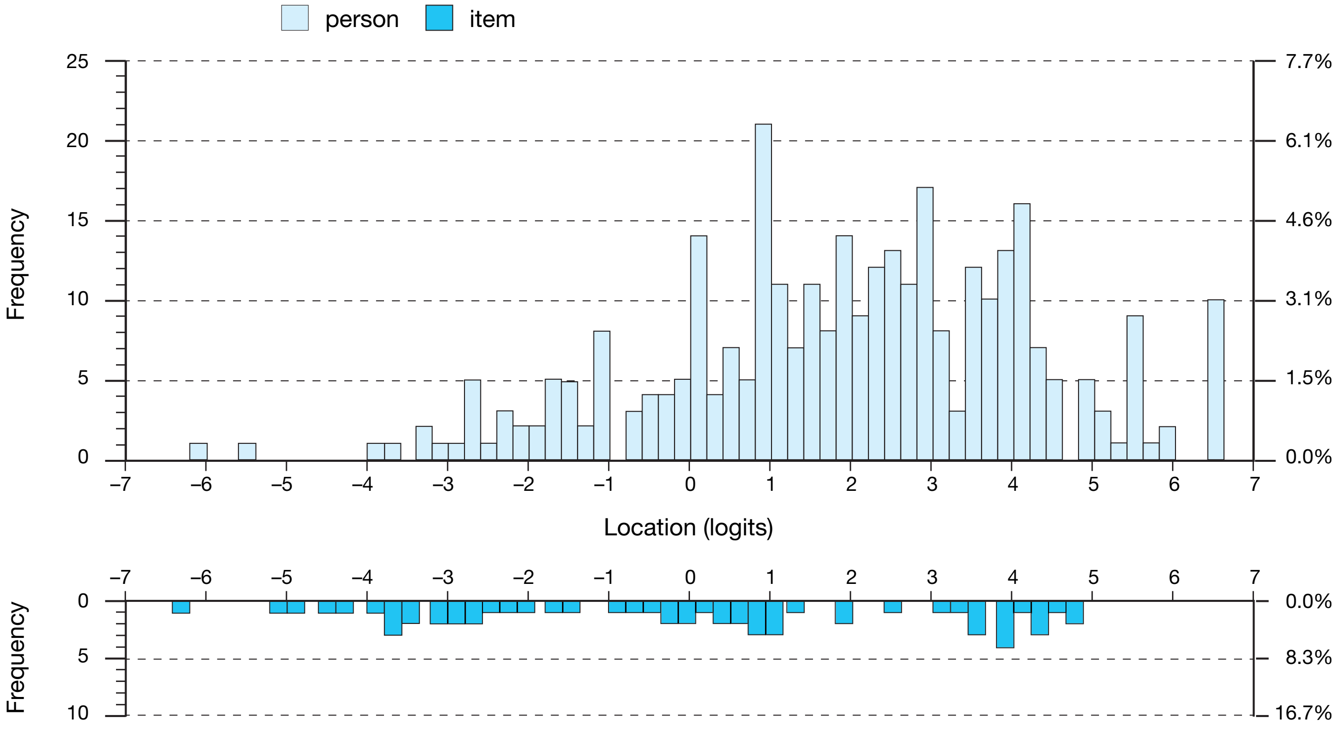

Figure 3. Person-item threshold distribution for Sample 1 (n = 326). Grouping set to interval length of 0.20 making 70 groups.

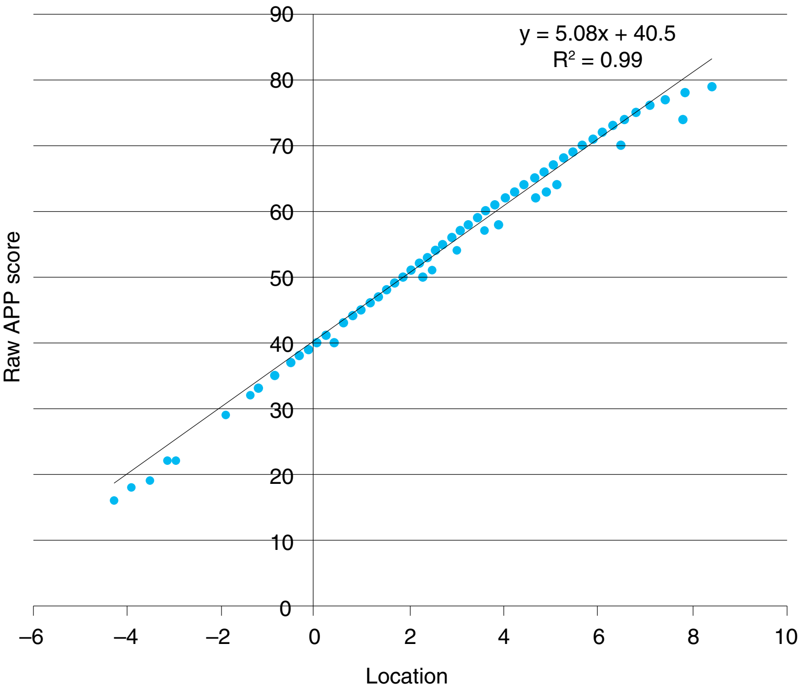

Figure 4. Plot of person logit location and raw APP score (Sample 1 n = 326).

differential item functioning was demonstrated in either of the two samples for the following variables: the student's age, gender, or amount of prior clinical experience, the educator's age, gender, or experience as an educator, or the type of facility, university, or clinical area. This indicates the APP item ratings were not systematically affected by any of these nine variables.

Local independence and dimensionality: Local independence is the assumption that responses to items are independent. 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 Item 1 (Demonstrates an understanding of patient rights and consent), Item 2 (Demonstrates a commitment to learning), Item 3 (Demonstrates ethical, legal and culturally sensitive practice), and Item 5 (Verbal communication).

A unidimensional set of items measures a single underlying construct. APP dimensionality was tested by an independent t-test procedure of person ability locations derived from two subsets of items – one loading positively and the other negatively > 0.30 on the first residual factor of the principal components analysis in RUM2020 (Tennant and Pallant 2006). The proportion of persons with significantly different person estimates based on the two item subsets was 7.3% and 6.9% for the two samples. The confidence intervals for a binomial test of proportions both included 5%, providing evidence of the unidimensionality of the scale.

Figure 4 shows the relationship between raw ordinal APP scores and person logit location for Sample 1. Sample 2 exhibited the same relationship.

Discussion

This second and final field trial of the 20-item APP confirmed that it is a unidimensional instrument with a response scale that is used as anticipated and that is able to discriminate at least four distinct levels of student performance.

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, such as communication and professional behaviours, and those that are more difficult and therefore may be expected to take longer to master. The hierarchies of both samples in the current study revealed that items related to analysis and planning (critical thinking), goal setting, and selection and progression of interventions were the most difficult items for students to perform.

Rheault and Coulson (1991) demonstrated a similar ranking of a 6-item physiotherapy practice assessment instrument. 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 the data collected in the field test demonstrated overall fit to the Rasch model for both participant samples, Item 6 (Written communication) showed misfit to the Rasch model. Pallant and Tennant (2007) state that one of the most common sources of item misfit is respondents' (educators) inconsistent use of the scoring options resulting in disordered thresholds. However, investigation of threshold ordering of the 20 polytomous items on the APP showed there were no disordered thresholds in either sample. Despite a small improvement in overall item fit to the model when Item 6 was deleted, removal of this item is not justified given that written communication is part of the current Australian Standards for Physiotherapy (Australian Physiotherapy Council 2006) and represents an essential aspect of professional competence. Exploration of this issue with clinical educators suggests that there is a lack of consensus with respect to the timing of recording patienttherapist interactions during or after the encounter, and that agencies did not clearly communicate their expectations to students early in the placement. Further research on this item and how it is being interpreted and scored by educators is warranted.

In the final field test no significant differential item functioning 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 that 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.

One of the primary advantages of Rasch analysis is that raw ordinal scores may be converted to interval level Rasch scores. Given the almost perfect linear relationship between Rasch logit scores and raw scores shown in Figure 4, the complexity associated with converting the raw score to a Rasch score does not appear warranted.

The APP was developed collaboratively, tested within the constraints of a dynamic and unpredictable clinical environment, and has been taken up almost universally as the assessment instrument in entry-level physiotherapy programs in Australia and New Zealand. The advantages of a single, national instrument are the reduction of assessment burden on clinical educators dealing with students from multiple university programs, and the standardardisation of student assessment for entry-level practice ensuring that students are assessed against the same performance indicators, on the same rating scale, against explicit standards for entry-level practice.

The evidence of construct validity provided by Rasch analysis supports the interpretation that a student's score on the APP is an indication of their underlying level of professional competence as demonstrated during workplace-based placements. The reliability of judgements made with the APP will be published separately. ■

eAddenda: Figure 1 and Appendix 1 available at jop.physiotherapy.asn.au

Ethics: Approval for the study was provided by the Human Ethics Committees of the nine participating universities. All participants gave written informed consent before data collection began.

Support: Funding from the Australian Learning and Teaching Council (ALTC) enabled employment of a research assistant and travel to conduct focus groups and training workshops.

Acknowledgements: Thanks go to the clinical educators and students who participated, to the University Clinical Education MAnagers of Australia and New Zealand, and to the Council of Physiotherapy Deans, Australia and New Zealand, who championed the development of a national assessment instrument.

Correspondence: Dr Megan Dalton, Department of Physiotherapy, School of Primary Health Care, Monash University, Australia. Email: megan.dalton@monash.edu

References

Andrich D, Lyne A, Sheridan B, Luo G (2003) RUMM2020. *RUMM Laboratories* Perth.

Australian Physiotherapy Council (2006) http://www.physiocouncil.com.au/australian_standards_for_physiotherapy/. [Accessed 16/08/2010].

Bond TG, Fox MT (2007) Applying the Rasch Model. Fundamental measurement in the human sciences (2nd edn). Mahwah, NJ: Erlbaum.

Cliff N, Keats JA (2003) Ordinal measurement in the behavioral sciences. Mahwah, NJ: Erlbaum.

Coghlan D, Brannick T (2001) Doing action research in your own organisation. Thousand Oaks, CA: Sage.

Dalton M, Keating J, Davidson M (2009) 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 pp 6–28]. Brisbane: Griffith University. Available online at: www.altc.edu.au

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: 283–294.

Pallant JF, 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: 1–18.

Prescott-Clements L, van der Vleuten CP, Schuwirth LW, Hurst Y, Rennie JS (2008) Evidence for validity within workplace assessment: the Longitudinal Evaluation of Performance (LEP). *Medical Education* 42: 488–495.

Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Rethans JJ, Norcini JJ, Baron-Maldonado M, Blackmore D, Jolly BC, LaDuca T et al (2002) The relationship between competence and performance: implications for assessing practice performance. *Medical Education* 36: 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: 10–13.

Southgate L, Hays RB, Norcini J, Mulholland H, Ayers B, Woolliscroft J et al (2001) Setting performance standards for medical practice: a theoretical framework. *Medical Education* 35: 474–481.

Streiner DL, Norman GR (2003) Health Measurement Scales. A practical guide to their development and use (3rd edn). New York: Oxford University Press.

Tennant A, Conaghan PG (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: 1358–1362.

Tennant A, Pallant JF (2006) Unidimensionality matters! (a tale of two Smiths). *Rasch Measurement Transactions* 20: 1048–1051.

van der Vleuten C (2000) Validity of final examinations in undergraduate medical training. *British Medical Journal* 321: 1217–1219.

Wilson M (2005) Constructing measures: an item response modeling approach. Mahwah, NJ: Erlbaum.

Website

Dalton MB (2011) Development of the Assessment of Physiotherapy Practice - A standardised and validated approach to assessment of professional competence in physiotherapy. Doctor of Philosophy Thesis, Monash University, Melbourne. URL: http://arrow.monash.edu.au/hdl/1959.1/479140