The Assessment of Physiotherapy Practice (APP) is a reliable measure of professional competence of physiotherapy students: a reliability study

Megan Dalton[1,3], Megan Davidson[2] and Jennifer L Keating[1]

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

Question: What is the inter-rater reliability of the Assessment of Physiotherapy Practice (APP) instrument, and what is the error associated with individual scores? **Design**: Cross-sectional inter-rater reliability study. Thirty pairs of clinical educators each assessed one student after observing student practice over a 5-week clinical placement. **Participants**: Sixty clinical educators from five Australian universities formed 30 independent pairs of assessors. **Outcome measures**: Each pair completed two independent assessments of one student, providing 60 completed APP assessments and an associated Global Rating Scale score for analysis. **Analysis**: Correlational coefficients and measurement error expressed in APP scale units were computed to provide a comprehensive analysis of the likely utility of APP scores and to enable score and change score interpretation. **Results**: Percentage of agreement between assessors for each item ranged from 56% (Item 19, evidence-based practice) to 83% (Item 20, risk management) and across all items averaged 70% (SD 7). The ICC(2,1) was 0.92 (95% CI 0.84 to 0.96) for the total APP score and 0.72 (95% CI 0.50 to 0.86) for the Global Rating Scale. The standard error of measurement for the total score (scale width 0–80) was 3.2 APP points and the MDC_{90} was 7.86 representing 9% of the scale width. Bland-Altman analyses identified no systematic differences between raters. **Conclusion**: Clinical educators demonstrated a high level of reliability when using the APP instrument to assess physiotherapy students' level of professional competence in workplace-based practice. **[Dalton M, Davidson M, Keating JL (2012) The Assessment of Physiotherapy Practice (APP) is a reliable measure of professional competence of physiotherapy students: a reliability study. *Journal of Physiotherapy* 58: 49–56]**

Key words: Educational measurement, Professional competence, Clinical competence, Physical therapy (Specialty), inter-rater reliability, intraclass correlation coefficient, Physiotherapy

Introduction

The Assessment of Physiotherapy Practice (APP) is a 20-item instrument covering professional behaviour, communication, assessment, analysis and planning, intervention, evidence-based practice, and risk management. Each item is assessed on a 5-level scale from 0 (Infrequently/rarely demonstrates performance indicators) to 4 (Demonstrates most performance indicators to an excellent standard). A rating of 2 (Demonstrates most performance indicators to an adequate standard) indicates that the minimum standard for an entry-level physiotherapist has been met. The total APP score ranges from 0 to 80. Rasch analysis of APP scores indicated that the 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 (Dalton et al 2011). The APP has been widely adopted by entry-level physiotherapy programs in Australia and New Zealand.

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 (Baartman et al 2007, Epstein and Hundert 2002, Roberts et al 2006). 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. In the context of assessment of workplace performance, reliability is the extent to which assessment yields relatively consistent results across occasions, contexts and assessors (Baartman et al 2007). Reliability is dependent on the characteristics of the test, the conditions of administration, the group of examinees and the interaction between these factors (Streiner and Norman 2003, Wolfe and Smith 2007). While repeated, blinded testing of the same student under the same conditions in the authentic practice environment by the same assessor is not feasible in performance-based assessment, the consistency with which different assessors rate the performance of different students (inter-rater reliability) is achievable. Since inter-rater reliability

What is already known on this topic: The Assessment of Physiotherapy Practice (APP) is a valid measure of the clinical competence of physiotherapy students. It covers professional behaviour, communication, assessment, analysis, planning, intervention, evidence-based practice and risk management.

What this study adds: Clinical educators demonstrate a high level of reliability using the APP to assess students in workplace-based practice.

Table 1. Participant and placement characteristics.

Characteristic	University 1	University 2	University 3	University 4	University 5
Program	4-year bachelor degree	4-year bachelor degree	4-year bachelor degree	4-year bachelor degree	5-year double degree
Year of study	3	3	3/4	3	5
Students, n male:female	1:3	3:3	2:4	3:2	3:6
Student age (yr), mean (SD)	22 (3)	22 (3)	22 (3)	23 (3)	23 (3)
Clinical educators, n male:female	3:5	4:8	5:7	4:6	6:12
Clinical educator age (yr), mean (SD)	39 (9)	37 (8)	33 (9)	36 (9)	35 (9)
Facility type	Hospital	Hospital	Hospital	Hospital	Hospital
Clinical area/s	Orthopaedics (inpatients), Musculoskeletal (outpatients)	Cardiorespiratory, Paediatrics	Neurological rehabilitation, Community health	Cardiorespiratory, Gerontology rehabilitation	Orthopaedics (inpatients), Musculoskeletal (outpatients), Paediatrics

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 and 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 address them with methods such as standardised training of assessors, or adjustment of grades for areas of practice where higher level skills are typically expected (eg, critical care wards). Random errors are, by their nature, unpredictable. They need to be estimated and allowed for in score interpretation (Rankin and Stokes 1998).

The research question was therefore:

> What is the inter-rater reliability of the APP instrument, and what is the error around individual scores?

Method

This reliability study was conducted in the authentic practice environment to investigate the error in APP measurements in the typical application of the instrument (Baartman et al 2006).

Design

The inter-rater reliability trial was a cross-sectional study designed to replicate authentic assessment procedures. Sixty clinical educators formed 30 independent pairs of assessors. Since not all physiotherapy education programs typically utilised shared supervision (ie, two supervisors sharing supervision of a student), five programs where this routinely occurred were identified from the twelve physiotherapy entry-level programs in Australia and clinical educators were invited to participate in the trial.

Replication of authentic practice meant that the assessors provided educational supervision to the students during the clinical placement and then each student (n = 30) was assessed independently by their unique pair of educators using the APP at the end of a five-week clinical placement block. The blocks were scheduled across one university semester. Educators completed the APP and also gave students a rating of overall performance, on a Global Rating Scale of *not adequate*, *adequate*, *good*, or *excellent*. Students, working with supervision, provided physiotherapy services during this placement on a full-time basis (32–40 hours/week). Approval for the study was obtained from the human ethics committees of each of the five participating universities.

Participants

Students enrolled in entry-level physiotherapy programs from five universities in Australia were assessed by educators using the APP on completion of a five-week full-time clinical placement block. 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). 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.

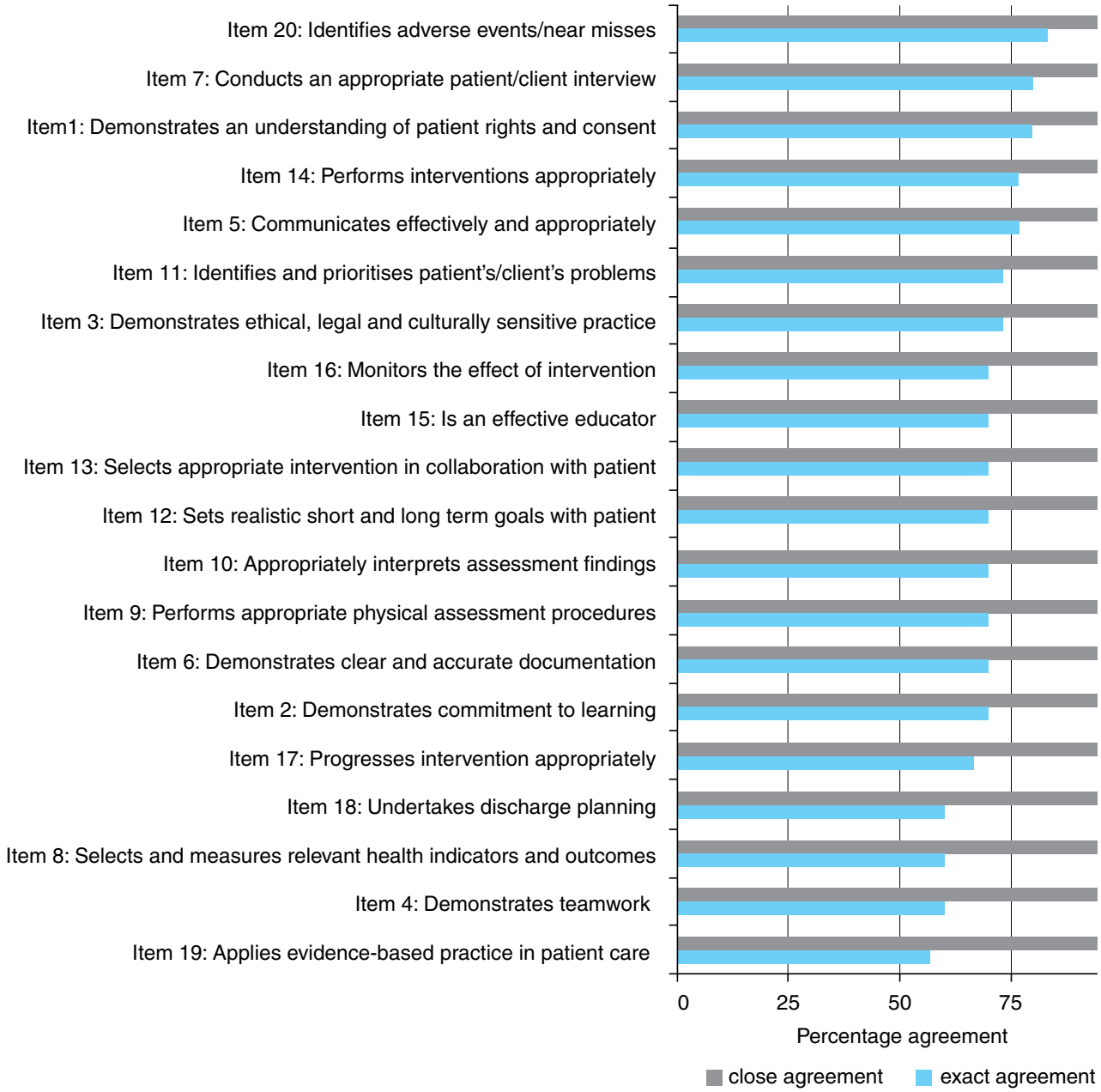

Figure 1. Percentage agreement between raters for 20 items on APP. Percent close agreement is within 1 point on the 5-point scale.

Inter-rater reliability trial procedure

Information on the reliability trial was provided in writing to the educators and students and their written consent to participation was obtained. All clinical educators received training in the use of the APP through workshop attendance and/or access to the APP resource manual. 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. To be eligible to participate, each pair of educators 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. Assessment data were excluded from analysis if either the student or their clinical educator did not consent to participate in the research and if any pair of assessors did not complete the APP instrument as per the instructions that both assessors must complete the APP independently within 12 hours of each other. Participants were advised that all data would be permanently de-identified prior to data analysis.

Data management and analysis

On completion of each placement the completed APP forms were returned by mail; data were entered into a spreadsheet, matched to the paired report, and de-identified prior to analysis. Planned data analysis included: descriptive statistics; calculation of Pearson's r and the Intraclass Correlation Coefficient (ICC 2,1) (two-way random-effects model) (and their confidence intervals), the standard error of

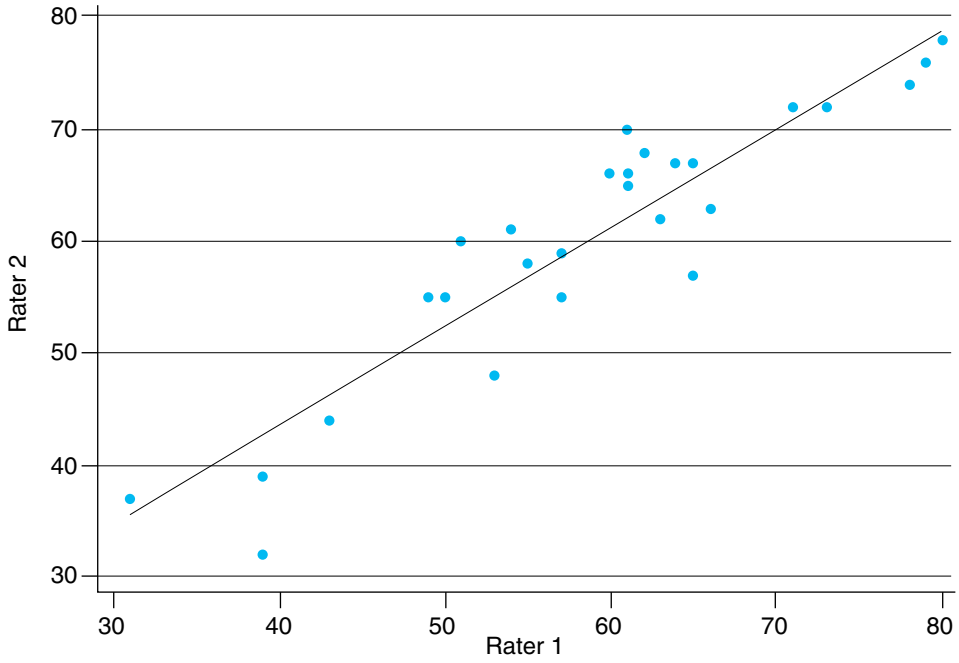

Figure 2. Scatterplot of APP scores for Rater 1 and Rater 2.

measurement (SEM) and the minimum detectable change at 90% confidence (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 and 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 Global Rating Scale scores was calculated.

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 and Wallnau 2005). The research team considered that 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.

Results

Thirty-three pairs of clinical educators (66 independent educators) and 33 independent third and fourth year physiotherapy students consented to participate in the reliability trial. 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. Table 1 presents a summary of participant characteristics.

Percentage agreement between raters

Ratings by two assessors for 14 of the 20 APP items were identical among 70% or more of the 30 pairs. Figure 1 shows the percent exact agreement and the percent close agreement, ie, within 1 point on the 5-point scale, for each of the 20 items.

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 4-point Global Rating Scale.

Pearson's product-moment correlation coefficient

A scatterplot was visually assessed for violation of assumptions of linearity and homoscedasticity. Figure 2 shows the positive, strong (Cohen 1988), linear, significant relationship between Rater 1 and Rater 2 total APP scores [$r = 0.92$ (95% CI 0.87 to 0.95), $p < 0.0005$]. The coefficient of determination ($r^2 = 0.85$) indicates that 85% (95% CI 75% to 90%) of the variance in a rater's scores was explained by variance in the other rater's scores.

Intraclass Correlation Coefficient (ICC(2,1))

The ICC(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). The ICC(2,1) for the global rating scale scores was 0.72 (95% CI 0.50 to 0.86). Table 2 presents the ICC(2,1) results for the total score, each of the 20 APP items, and the Global Rating Scale.

Standard Error of Measurement (SEM)

The SEM 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 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.

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

Table 2. Intraclass correlation coefficient (ICC), standard error of the measurement (SEM) and minimum detectable change (MDC_{90}) for the total APP score, global rating scale, and individual APP items.

	ICC(2,1)[a]	95% CI	2SEM	MDC_{90}
Total APP score	0.92	0.84 to 0.96	60.5	70.86
Global Rating Scale	0.72	0.50 to 0.86	0.84	0.98
Professional behaviour				
Item 1: Demonstrates an understanding of patient/client rights and consent	0.81	0.64 to 0.90	0.31	0.69
Item 2: Demonstrates commitment to learning	0.70	0.46 to 0.85	0.35	0.70
Item 3: Demonstrates ethical, legal and culturally sensitive practice	0.77	0.57 to 0.88	0.35	0.77
Item 4: Demonstrates teamwork	0.65	0.37 to 0.81	0.45	0.64
Communication				
Item 5: Communicates effectively and appropriately – verbal/non-verbal	0.82	0.66 to 0.91	0.30	0.85
Item 6: Demonstrates clear and accurate documentation	0.79	0.56 to 0.89	0.31	0.80
Assessment				
Item 7: Conducts an appropriate patient/client interview	0.80	0.62 to 0.90	0.30	0.80
Item 8: Selects and measures relevant health indicators and outcomes	0.60	0.29 to 0.77	0.43	0.61
Item 9: Performs appropriate physical assessment procedures	0.71	0.48 to 0.85	0.38	0.71
Analysis and planning				
Item 10: Appropriately interprets assessment findings	0.63	0.35 to 0.80	0.37	0.65
Item 11: Identifies and prioritises patient's/client's problems	0.75	0.53 to 0.87	0.36	0.74
Item 12: Sets realistic short and long term goals with the patient/client	0.76	0.55 to 0.87	0.35	0.75
Item 13: Selects appropriate intervention in collaboration with patient/client	0.73	0.50 to 0.86	0.35	0.73
Intervention				
Item 14: Performs interventions appropriately	0.82	0.66 to 0.91	0.29	0.85
Item 15: Is an effective educator	0.82	0.65 to 0.90	0.35	0.81
Item 16: Monitors the effect of intervention	0.60	0.32 to 0.79	0.38	0.60
Item 17: Progresses intervention appropriately	0.76	0.57 to 0.88	0.36	0.77
Item 18: Undertakes discharge planning	0.71	0.49 to 0.85	0.44	0.71
Evidence-based practice				
Item 19: Applies evidence based practice in patient care	0.70	0.43 to 0.83	0.44	0.68
Risk management				
Item 20: Identifies adverse events/near misses and minimises risk associated with assessment and interventions	0.74	0.52 to 0.86	0.34	0.75

[a] = all ICC $p < 0.0005$

(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 to 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) indicates that real change in performance on that item has occurred beyond random variability.

Bland-Altman analyses

A Bland and Altman plot was constructed to display errors in estimates of total APP scores (Figure 3). 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. The Bland-Altman plot shows that the disagreement between raters was not greater among high scores than among low scores, or vice versa. Errors 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.

Discussion

In this inter-rater reliability study of APP scores, the percentage agreement for individual items 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 Global Rating Scale.

Individual item ICCs ranged from 0.60 for Item 8 (selecting relevant health indicators and outcomes) and Item 16 (monitoring the effect of intervention), to 0.82 for Item 5

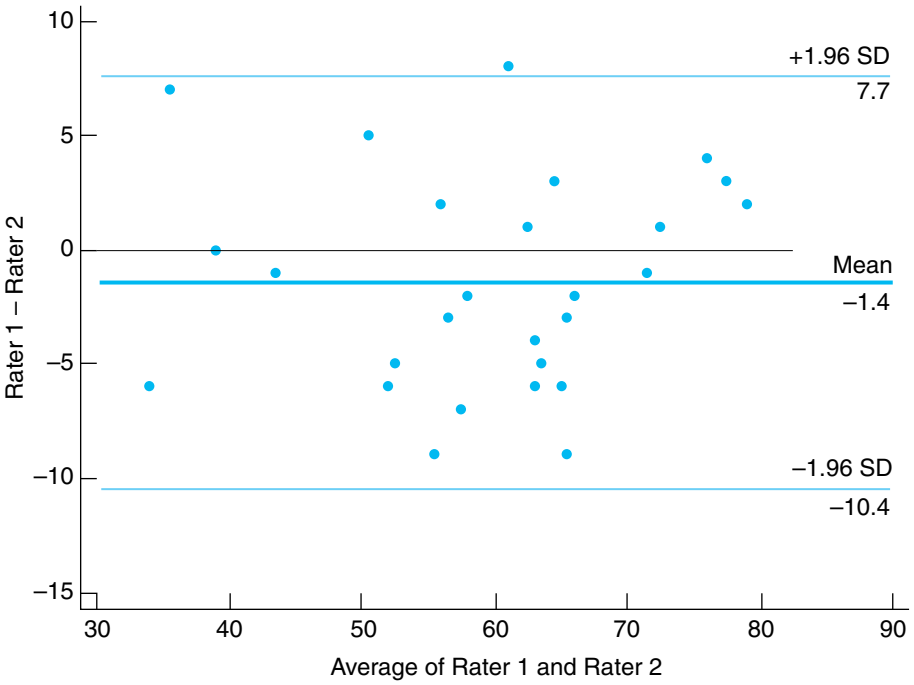

Figure 3. Plot of the differences between raters' marks against the means of raters' marks for the total score out of 80 (n = 60 assessments). The mean difference between raters bisects the y-axis and the upper and lower lines represent the 95% limits of agreement.

(verbal communication), Item 14 (performing interventions), and Item 15 (being an effective educator). The ICC(2,1) for total APP scores for the two raters was 0.92 (95% CI 0.84 to 0.96), while the SEM of 3.2 and MDC_{90} of 7.86 allows scores for individual students to be interpreted relative to error in the measurement.

It should be noted that while 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 and Farr 1980). Other studies suggest that rater behaviour may contribute less to error variance than other factors such as student knowledge, tasks sampled, and case specificity (Govaerts et al 2002, Keen et al 2003, Shavelson et al 1993).

A limitation of the current study is that while the paired assessors were instructed not to discuss the grading of student performance during the five-week clinical placements, adherence to these instructions was not assessed. 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. This may have reduced the independence of the rating given by the paired raters, and inflated the correlation coefficient. Mitigating this was that, in all 30 pairs of raters, the education of students was shared with little, if any, overlap of work time between raters. While this trial design limited opportunities for discussion between raters, educators who regularly work together or job share a position may be more likely to agree even if there is little, if any, overlap in their work time. Further research investigating the influence a regular working relationship may confer on assessment outcomes is required.

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 practice was given priority and the possible limitations relating to this method were considered acceptable.

Four studies have investigated inter-rater reliability of physiotherapy clinical performance assessment instruments. Intraclass correlations (2,1) of 0.87 for the total Clinical Performance Instrument (CPI) score were found for joint evaluators of physiotherapy students and 0.77 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 for the Common Assessment Form (CAF), and Meldrum et al (2008) reported an ICC of 0.84 for a predecessor to the CAF. Loomis (1985) reported ICCs of 0.62 and 0.59 for third and fourth year total scores respectively on the Evaluation of Clinical Competence form.

A range of expressions of test reliability have been provided in this study. 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 0–80 scale width. The reliability of the APP compares favourably with reliability estimates reported by others who have developed instruments for assessing competency to practise physiotherapy. Coote et al (2007) and Meldrum et al (2008) reported data that enabled calculation of the SEM and it appears that for the Common Assessment Form and its predecessor this was also 3% to

4% on a 0–80 scale. The evidence suggests that clinicians are reasonably consistent in their judgements of student ability to practise 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 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. 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 the test/retest correlation of 0.92. Additionally in other data that we have collected (Dalton 2011), students commencing workplace-based 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. Across the practice period 77% of students change 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.

However, it 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. 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 indicated 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.

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 the APP scores behaved as quality of life scores do, then an estimate of the possible minimally important difference would be 6–8 points, a proposal that warrants investigation.

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 consistency in the assessment and marking of physiotherapy students' performance on clinical placements when using the Assessment of Physiotherapy Practice. ■

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

Competing interests: Nil.

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: The authors acknowledge the assistance of Curtin, James Cook, La Trobe, Griffith, Monash, and Sydney Universities and thank the clinical educators and students who participated.

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

References

Alexander HA (1996) Physiotherapy student clinical education: the influence of subjective judgements on observational assessment. *Assessment & Evaluation in Higher Education* 21: 357.

Baartman LKJ, Bastiaens TJ, Kirschner PA, van der Vleuten CPM (2006) The wheel of competency assessment: presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation* 32: 153–170.

Baartman LKJ, Bastiaens TJ, Kirschner PA, van der Vleuten CPM (2007) Evaluating assessment quality in competence-based education: a qualitative comparison of two frameworks. *Educational Research Review* 2: 114–129.

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1: 307–310.

Cohen JW (1988) Statistical power for the behavioral sciences (2nd ed). Hillsdale: Lawrence Erlbaum.

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: 6–10.

Dalton M, Davidson M, Keating JL (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.

Epstein RM, Hundert EM (2002) Defining and assessing professional competence. *Journal of American Medical Association* 287: 226–235.

Govaerts MJ, van der Vleuten CP, Schuwirth LW (2002) Optimising the reproducibility of a performance-based assessment test in midwifery education. *Advances in Health Science Education* 7: 133–145.

Gravetter F, Wallnau L (2005) Essentials of statistics for the behavioural sciences. Pacific Grove: Wadsworth.

Keen AJ, Klein S, Alexander DA (2003) Assessing the communication skills of doctors in training: reliability and sources of error. *Advances in Health Sciences Education* 8: 5–16.

Landy FJ, Farr JL (1980) Performance rating. *Psychological Bulletin* 87: 72–107.

Loomis J (1985) Evaluating clinical competence of physical therapy students. Part 2: assessing the reliability, validity and usability of a new instrument. *Physiotherapy Canada* 37: 91–98.

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.

Norman GR, Sloan JA, Wyrwich KW (2003) Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care* 41: 582–592.

Rankin G, Stokes M (1998) Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clinical Rehabilitation* 12: 187–199.

Roberts C, Newble D, Jolly B, Reed M, Hampton K (2006) Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher* 28: 535–543.

Shavelson RJ, Gao X, Baxter G (1993) Sampling variability in performance assessments. CSE Technical report number 361. Santa Barbara: National Center for Research on Evaluation, Standards and Student Testing, University of California.

Streiner DL, Norman GR (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: 329–353.

Wilson M (2005) Constructing Measures: An item response modeling approach. Mahwah, New Jersey: Lawrence Erlbaum.

Wolfe EW, Smith EV Jr (2007) Instrument development tools and activities for measure validation using Rasch models: Part I – instrument development tools. *Journal of Applied Measurement* 8: 97–123.

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. http://arrow.monash.edu.au/hdl/1959.1/479140

Statement regarding registration of clinical trials from the Editorial Board of *Journal of Physiotherapy*

All clinical trials submitted to *Journal of Physiotherapy* for publication must have been registered in a publicly-accessible trials register. We will accept any register that satisfies the International Committee of Medical Journal Editors requirements. Authors must provide the name and address of the register and the trial registration number on submission. Trials that have been registered prospectively will be given higher priority. From 2013 the journal will only accept trials that have been registered prospectively unless data collection began before 2006, in which case retrospective registration is acceptable.

