EVALUATION OF A NEW CLINICAL PERFORMANCE ASSESSMENT TOOL: A RELIABILITY STUDY

Joseph C (MSc)[1]
Frantz J (PhD)[1]
Hendricks C (MSc)[2]
Smith M (PhD)[1]

[1] University of the Western Cape,
[2] University of Cape Town.

ABSTRACT: *Clinical practice is an essential requirement of any graduate physiotherapy programme. For this purpose, valid and reliable assessment tools are paramount for the measurement of key competencies in the real-world setting. This study aims to determine the internal consistency and inter-rater reliability of a newly developed and validated clinical performance assessment form. A cross-sectional quantitative research design was used, which included paired evaluations of 32 (17 treatment and 15 assessment) student examinations performed by two independent clinical educators. Chronbachs alpha was computed to assess internal consistency and intraclass correlation coefficient (ICC's) with confidence intervals of 95% were computed to determine the percentage agreement between paired examiners. The degree of internal consistency was substantial for all key performance areas of both examinations, except for time and organisational management (0.21) and professionalism (0.42) in the treatment and evaluation examinations respectively. The overall internal consistency was 0.89 and 0.73 for both treatment and assessment examinations, indicating substantial agreement. With regard to agreement between raters, the ICC's for the overall marks were 0.90 and 0.97 for both treatment and assessment examinations. Clinical educators demonstrated a high level of reliability in the assessment of students' competence using the newly developed clinical performance assessment form. These findings greatly underscore the reliability of results obtained through observation of student examinations, and add another tool to the basket of ensuring quality assurance in physiotherapy clinical practice assessment.*

KEY WORDS: *CLINICAL PERFORMANCE ASSESSMENT FORM, INTERNAL CONSISTENCY, INTER-RATER RELIABILITY.*

INTRODUCTION

Health professional programmes have a core goal of ensuring that graduates are fit for practice in their chosen disciplines. Determining the level of competency of the student is a complex process that occurs in the context of various clinical settings, various types of patients, and the expertise and experience of the clinical educator (Alexander 1996). Clinical practice is an essential requirement of any graduate physiotherapy curriculum worldwide, thus relying on the objective evaluation of the students' clinical performance before they could be promoted to the next year level or graduate from the programme. According to Strohschein et al (2002), clinical practice provides students with the platform to put theoretical knowledge into practice and create the opportunity to fine-tune competencies that are important for the holistic management of the patients. However, the reliable evaluation of these competencies by the clinical educators is a challenge and attempts at standardising this process are essential (Cross 2001). These challenges could be intrinsic to the clinical educator and closely related to the assessment tool utilised for the evaluation of students' performance. It has been found that clinical instructors determine students' abilities/competency from a general impression of their performance and considering the execution of skills and demonstration of behaviour as a whole, rather than distinctly (Cross & Hicks 1997; Alexander 1996). Furthermore, clinical educators tend to exhibit their subconscious bias and their previous learning experiences when examining students' clinical performance (Meldrum et al 2008).

In South Africa, students are expected to be evaluated on their competence within the clinical context by external examiners which are not affiliated to the home university. These procedural criteria for validating clinical competence add challenges to the examination process due to varying expectations, standards, procedures and key criteria used by external examiners (Meldrum et al 2008). In order to level the playground between the students wanting to achieve and examiners trying to be consistent throughout the process, standardised outcome measures are imperative.

In an attempt to standardise measurements of students' clinical performance, valid and reliable measurement tools should be developed for the assessment of clinical skills and behaviours (Jette et

Correspondence Author:
Conran Joseph
Department of Physiotherapy
University of Western Cape
Private Bag X17
Bellville 7530
South Africa
Email: cjoseph@uwc.ac.za

al 2007). The American Physical Therapy Association (1997) advocates that measuring tools should have similarities in terms of skills and student behaviours important for the evaluation and the expected outcome of the clinical practice module to align with the requirements of entry-level clinical performance. Moreover, discussion of issues in developing valid and reliable assessments must include consideration of the statistical strategies used to guide the design and the validation of the tool to ensure that it measures competency accurately. According to Wilkinson (2007, p635), '...the validity of assessments comes from using appropriate tools that appropriately sample from a curriculum blueprint'. In addition, reliability comes from aggregating observations from a variety of situations using a variety of assessors.'

In an attempt to respond to the criteria of standardising the evaluation process of students' clinical performance the authors of this paper developed a clinical performance tool, using the Delphi method, specifically for third and fourth year physiotherapy students (Joseph et al 2011). The Delphi technique was used to obtain an informed or refined consensus from a group of experts in the field of clinical education (Hassan et al 2000). The tool was developed based on the key performance areas and associated assessment criteria essential for the evaluation of students' performance in a South African physiotherapy training context. The development of the clinical performance assessment form required four rounds of comments and feedback from the Delphi panel and confirmation of interpretations by the task team in the subsequent rounds. Using experts in the field of study has been deemed acceptable for establishing face and content validity (Joseph et al 2011). The panel of experts used in this study found the newly-designed instrument to have face- and content validity as the inclusion of each key performance area and associated criteria was based on an agreement of at least 75% of the responses by the panel. In addition to validation, the instrument was also being tested for reliability. This study adopted a similar method of reliability testing as Meldrum et al (2008). Since students rotate through various clinical settings and speciality areas, this study aimed at determining the inter-rater reliability and internal consistency of the tool within various clinical contexts.

METHODS

Study design

This study investigated scale development and the psychometric properties of a measure aimed at assessing clinical competence. In particular, the reliability of the measure on a clinical sample was assessed by using quantitative methods. The cross-sectional design was deemed appropriate given that the study did not attempt to make temporal inferences. The study first attempted to determine the internal consistency of the tool and then assessed whether there was agreement between raters when applying or using the instrument. Below is a brief outline of the scale development process.

Outcome measure: The clinical performance assessment form

A new clinical evaluation form was developed using the Delphi technique (Joseph et al 2011). This process highlighted eight key performance areas (KPA) with relative weighting to the overall score that needed to be assessed as part of determining clinical competence. This recommendation was implemented and reflects in the differential weighting assigned to KPAs for the treatment and assessment clinical evaluation form. The treatment form consisted of the following KPA with differential weighting assigned to each: knowledge (20), communication (10), planning (10), patient management (25), clinical reasoning (15), professionalism (10), reflection (5) and time management (5). The assessment form consisted of the same KPA, however the weighting differs for patient management (20), clinical reasoning (20) and reflection (10). The maximum cumulative score that could be obtained on the clinical evaluation form was 100.

Procedure The study was conducted during mock examinations of both 3rd and 4th year physiotherapy students with a total of 17 students performing a treatment examination and a total of 15 students performing an assessment examination within the selected clinical settings using two examiners per examination. According to Bonett (2002), a sample of 15 is sufficient for a reliability study with an estimated ICC correlation of 0.9. The tool was independently completed by two assessors. The assessors were clinical supervisors of the students in the specific clinical setting and an external examiner, which in this case was an academic experienced in the evaluation of students performance in the clinical context. Prior to commencing the examinations, the assessors were allowed to discuss the breakdown of the weighting of each criterion relative to the key performance area that it fell under. However, no changes were made to the distribution of marks within each key performance area.

Thirty-two paired examinations of clinical competence in 32 third and fourth year students were conducted by clinical supervisors and lecturers from at least four clinical settings at all levels of health care (primary, secondary and tertiary). The average years of clinical experience among the examiners was 19 years and of supervision experience was 12 years. The health conditions assessed included dysfunction of the neuromusculoskeletal-, neurological- (in both acute and rehabilitation settings), orthopaedic- and respiratory system. The evaluation of the reliability of the assessment form was also investigated within the Intensive Care Unit. The population for the reliability was all third and fourth year physiotherapy students at the various clinical placements. However, the sample was conveniently selected to ensure a variety of specialty areas as described above and the location of clinical settings for both 3rd and 4th year students. In the training programme where the tool was tested, the clinical settings are arranged in blocks whereby the students do their clinical rotations for a period of 4-6 weeks. Thus, the clinical supervisors selected for the study were in direct contact with the students by assisting the students' clinical performance and competencies in the management of patients once a week.

Data analysis

Data were entered into SPSS and analyzed using descriptive and inferential statistics. Chronbachs alpha was computed to assess internal consistency i.e. reflect the extent to which items of a test, measure various aspects of the same characteristic and nothing else. It 'correlates' each individual question with the total and gives a value between 0 to 1. A value of 0 indicates that none of the questions correlate well with the total. Value of 1 indicates excellent correlation. Usually a chronbach's alpha value of more than 0.6 is considered as adequate (Domholdt, 2000). Streiner (2003) supports the importance of internal consistency as an estimate of reliability, but underscores that caution be applied to the interpretation since the number of items could influence the alpha value obtained. Intraclass Correlation Coefficients (ICC's) with confidence intervals of 95% (95% Cl) were computed to determine the percentage agreement between paired examiners on the 32 examinations i.e. inter-rater agreement. The ICC was deemed appropriate since it measures how much of the total variance of scores can be attributed to differences between subjects (Bravo & Potvin 1991) and when replicate measures have no time sequence.

Table 1: Internal consistency scores per key performance area for the treatment and assessment examinations.

KPA	Treatment	Assessment
	Chronbach's alpha	Chronbach's alpha
Knowledge	0.90	0.80
Communication	0.83	0.60
Planning	0.81	0.70
Patient management	0.74	0.70
Clinical reasoning	0.85	0.81
Professionalism	0.84	0.42
Reflection	0.78	0.70
Time management and organizational skills	0.21	0.80
Total:	0.89	0.73

Table 2: Inter-rater agreement differences per key performance area

KPA	ICC Treatment	Mean (SD): Range Treatment in %	ICC Assessment	Mean (SD): Range Assessment in %
Knowledge	0.84	5 (5.9): 0-14	0.95	4.4 (4.4): 0-15
Communication	0.83	3.5(6.1): 0-20	0.44	11.6 (12): 0-40
Planning	0.81	7.0 (5.7): 0-20	0.67	8.1 (8.6): 0-30
Patient management	0.74	7.5 (6.3): 0-20	0.85	4.9 (5.8): 0-22
Clinical reasoning	0.84	8.3(5.3): 0-20	0.91	5.7 (7.3): 0-20
Professionalism	0.84	4.7 (5.0): 0-10	0.42	11.5 (11.3): 0-40
Reflection	0.79	7.0 (6.6): 0-20	0.36	18.2 (16): 0-60
Time management	0.49	15 (17): 0-60	0.56	14.0 (19): 0-40
Total	0.90	3.7 (2.7): 0-9	0.97	2.5 (3.6): 0-9

RESULTS

In total, 32 students were evaluated simultaneously by two raters in treatment examinations (n=17) and assessment examinations (n=15) respectively. Examinations were conducted in the following speciality areas; orthopaedics (4), respiratory (2), intensive care (6), outpatient settings including musculoskeletal disorders such as low back pain (8), sports injuries (4), neurology (5), and inpatient neurology (3).

Cronbach's alpha was the measure used to determine the internal consistency. The results show that the internal consistency for the treatment evaluation form was high, exceeding 0.7 for almost all the key performance areas except for time and organizational management (0.21). The internal consistency for the assessment evaluation form was high (>0.7) for all the key performance areas except professionalism. (Table 1).

The mean difference for the overall scores on the treatment examinations between the raters ranged from 0-9 marks with a mean difference of 3.7 (SD=2.7). Similarly, for the assessment examinations the mean difference ranged from 0-9 marks with an average of 2.5 (SD=3.6). Figures 1 and 2 show the maximum differences between raters in the total scores of both the assessment and treatment forms:

The inter-rater reliability for each of the key performance areas in both forms was calculated by using the Intraclass Correlation Coefficients between pairs of raters (see Table 2). Of the 15 pairs of assessment marks, moderate (0.56) to good (0.95) agreement between raters was found on the assessment form for five key performance areas and the remaining coefficients showed poor (0.36-0.44) agreement between the rating pairs involved for the assessment examinations. However, the overall scores had excellent agreement. In the treatment form, of the 17 pairs of raters, seven of the eight key performance area showed good agreement.

DISCUSSION

The effectiveness of assessment tools are gauged by its internal consistency and the level of agreement between raters when using the tool. Both these aspects

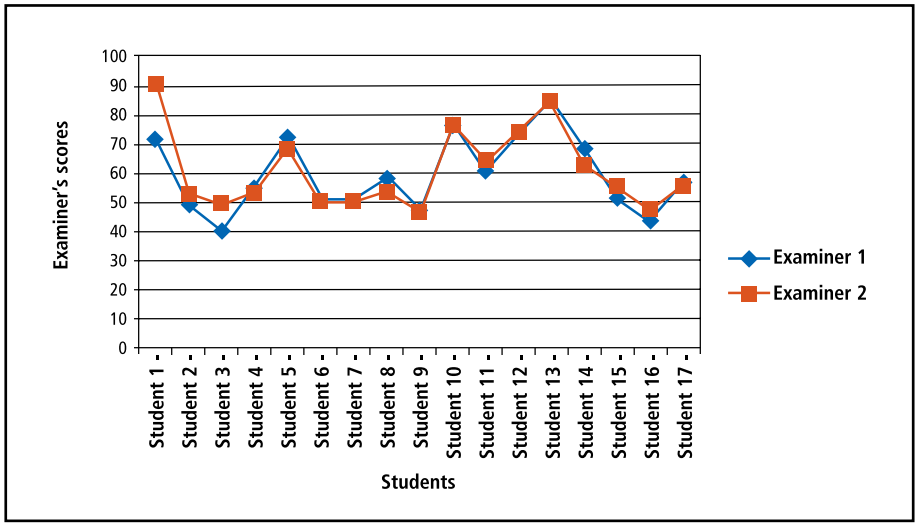

Figure 1: Treatment examination totals

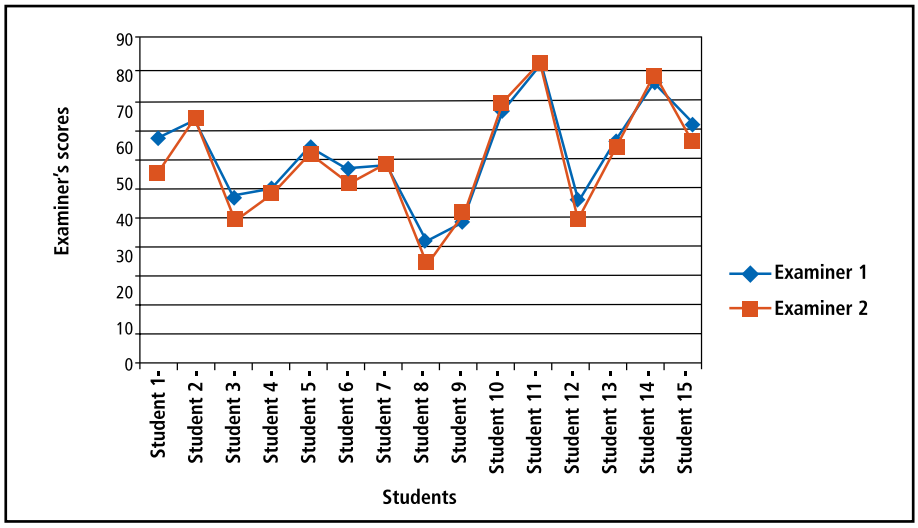

Figure 2: Examiner assessment scores

speak to the reliability of the measure as evidenced by its stability in terms of its construction (internal consistency) and inter-rater agreement. This article reports on the internal consistency and inter-rater reliability of a newly developed physiotherapy clinical assessment tool. The results of this initial study into the psychometric properties of the clinical assessment tool used for physiotherapy students in South Africa provide a baseline of the reliability of the measure.

The internal consistency, a measure of the extent to which individual items within the instrument fit the underlying construct, was found to exceed 0.70 for both the overall treatment (0.89) and the evaluation (0.73) form. The high degree of internal structure of both assessment forms serves as validation of the importance of each key performance area and assessment criteria in the holistic evaluation of students' clinical performance. However, the subscales related to organisational management and professionalism demonstrated weaker coefficients for internal consistency in the treatment and assessment examinations respectively. Despite the lower internal consistency of the aforementioned key performance areas, it was deemed prudent to retain these domains since their omission would have impacted negatively on the factor evaluation (Domholdt 2000), which was students' clinical performance in a South African healthcare context. A possible explanation for the reduced scores might be that there is too much discrepancy or variation in terms of how assessors interpreted the subscales. Thus it is recommended that the key performance areas related to time and organisational management, as well as professionalism be revised so that the variation in subjective interpretation is minimized. One possible solution is to ensure that the scoring key or guide for each criterion should become more directive and objective that in turn is likely to increase the overall internal consistency of the measure.

The findings indicated substantial agreement between raters in both treatment and assessment examinations as evidenced by intra-class correlations of 0.90 and 0.97 respectively. The level of agreement is considered substantial as per Bravo and Potvin's taxonomy for interpreting ICCs (1991) and even exceeds the ICCs (0.84) reported by Meldrum et al (2008). This study further found that where two assessors mark a student, the raters will be within 3.7 and 2.5 mark difference for a treatment and evaluation examination. This could have been as a result of the clear criteria for the various sections, examiners common understanding of the various sections and the years of experience of the examiners (12 years). This is confirmed by Meldrum et al (2008), who reported a difference of 6.2 marks between raters with supervision experience of about 2 years and therefore highlighted years of supervision experience as possible reason for discrepancies in mark allocations.

The higher agreement levels of the key performance areas relating to knowledge, clinical reasoning and patient management are indicative of the explicit nature with less arbitrary criteria in literature on the measurement of the ability of students to exhibit a wide-knowledge base specifically related to the patient's health condition, the identification of the problems encountered due to the health condition, which also underscores competence in clinical reasoning, and the ability to formulate and implement an appropriate management plan to address the holistic needs of the patient (Higgs et al 2005; Higgs & Jones 2000).

In the current study, the lowest mean level of agreement was found for the key area, time and organisational management (0.49). Other studies reported

that higher agreement levels (0.81) (Meldrum et al (2008) on this subscale are possible given that explicit criteria exist. It seems reasonable to conclude that the criteria for time and organisational management are ambiguous and lack explicit criteria on what constitutes competence in this domain. Another possible reason could be the importance examiners place on this domain, or the difference in the operationalisation of the concept and criteria of this domain. Time and organisational management is an important aspect in a South African Healthcare context (Joseph et al 2011) because of the high patients load and the lack of resources, therefore students are educated and formally assessed on this key area to ensure that they will cope once they graduate.

In the key performance areas labelled communication, professionalism and reflection, it became evident that assessors differed on the expectations of students and the degree of achievement. Guidelines on the definition, instillation and evaluation of professionalism abound in the literature on the education of health professionals, however the lack of consensus in this instance might be attributed to cultural differences and/or the context of professionalism. This poses considerable confusion for professional educators when referring to research on the evaluation of professionalism in students (van Mook et al 2009). Similarly, reflection is a complex task or process, thus relying on the conceptualisation of the clinical event and an objective approach to the process in order to deduce valuable learning experiences and identification of gaps in an individual's knowledge (Clouder 2000). The authors of this paper is of the opinion that educators go through the process of reflection based on the information gathered by the student during the assessment or treatment, thus painting a biased picture of the reflective process from their perspective and experience, and not the emerging clinical reasoning of the student. The format of the exam also does not allow sufficient time for students to underscore learning objectives, confirmation of the patient's problem and how it could be addressed.

CONCLUSION AND IMPLICATION

In conclusion, an instrument to assess clinical competency among physiotherapy students has been developed, validated and assessed for reliability. Overall the internal consistency and inter-rater reliability were substantial and acceptable. A few subscales could be made more explicit and training manuals should be made available with the core focus of improving reliability of all subscales. This study greatly improves the assessment practice and quality assurance procedure of such a controversial area in physiotherapy education. A next step would be to attempt to replicate this reliability study in other training programmes in South Africa.

REFERENCES

Alexander HA 1996 Physiotherapy student clinical education: the influence of subjective judgements in observational assessment. Assessment & Evaluation in Higher Education. 21: 357–367

American Physical Therapy Association 1997 Physical Therapist Clinical Performance Instrument. Alexandria, Va Bonett DG 2002 Sample size requirements for estimating intraclass correlations with desired precision. Statistics in Medicine *2*1 (9): 1331 -1335

Bravo G, Potvin L 1991 Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: toward the integration of two traditions. Journal of Clinical Epidemiology 44(4-5):381-390

Clouder L 2000 Reflective practice in physiotherapy education: a critical conversation. Studies in Higher education 25: 211-223

Cross V, Hicks C 1997 What do clinical educators look for in physiotherapy students? Physiotherapy 83: 249-260

Cross, V 2001 Approaching consensus in clinical competence assessment. Third round of a Delphistudy of academics' and clinicians' perceptions of physiotherapy undergraduates. Physiotherapy 87: 341–350

Domholdt E 2000 Physical Therapy Research: Principles and Application (2nd ed.). Philadelphia: WB Saunders Publishers

Hassan F, Keeney S, McKenna H 2000 Research guidelines for the Delphi survey technique. Journal of Advanced Nursing 32: 1008-1015

Higgs J, Jones MA 2000 Introduction: Clinical reasoning in the health professions. In J. Higgs & M. A Jones (Eds). Clinical reasoning in the health professions (2nd ed., pp. 3-32). Oxford: Butterworth-Heinemann.

Higgs J, McAllister L, Sefton, A 2005 Introduction: Communicating in the Health and Social Sciences. In J. Higgs, A. Sefton, A Street, L McAllister & I. Hay (Eds), Communicating in the health and social sciences (pp. 3-12). Oxford: University Press.

Jette DU, Bertoni, A, Coots R, Johnson H, Mclaughlin C, Weisbach C 2007 Clinical instructors perceptions of behaviours that comprise entry-level clinical performance in physical therapist student: A qualitative study. Physical Therapy 87: 833-843

Joseph C, Frantz J, Hendricks C 2011 Exploring the key performance areas and assessment criteria for the evaluation of students' clinical performance: A Delphi study. South African Journal of Physiotherapy 67 (2): 9-15

Meldrum D, Lydon A, Loughnane M, Geary F, Shanley L, Sayers K, shinnik E, Filan D 2008 Assessment of undergraduate physiotherapy clinical performance: Investigation of educator interrater reliability. Physiotherapy 94: 212-219

Streiner DL 2003 Starting at the beginning: an introduction to coefficient alpha and internal consistency. Journal of Personality Assessment 80: 99-103

Strohschein J, Hagler P, May L 2002 Assessing the need for change in clinical education practices. Physical Therapy 82:160–172

Van Mook WNKA, Gorter SL, de Grave WS, van Juijk SJ, O'Sullivan H, Zwaveling JH et al 2009 Professionalism beyond medical school: An educational continuum? European Journal of Internal Medicine 20(8): 148-152

Wilkinson T 2007 Assessment of clinical performance: gathering evidence. Internal Medicine Journal. 37(9): 631-636, p635