INTERNATIONAL JOURNAL OF HEALTH PROFESSIONS
Volume 4, Issue 1, 2017, Pages 53–65, ISSN 2296-990X, DOI: 10.1515/ijhp-2017-0008

DE GRUYTER OPEN

A systematic review of assessments for procedural skills in physiotherapy education

Assessment von prozeduralen Fähigkeiten in der physiotherapeutischen Ausbildung: Ein systematischer Review

Martin Sattelmayer[1,2]*, Roger Hilfiker[1], Gillian Baer[2]

[1]*School of Health Sciences, University of Applied Science and Arts Western Switzerland Valais (HES-SO Valais-Wallis), Leukerbad, Switzerland*
**martin.sattelmayer@gmail.com*

[2]*Queen Margaret University, School of Health Sciences, Physiotherapy, Edinburgh, Scotland*

Received 7 August 2016, accepted 27 February 2017, available online 5 May 2017

Abstract

Introduction: Learning of procedural skills is important in the education of physiotherapists. It is the aim of physiotherapy degree programmes that graduates are able to practice selected procedures safely and efficiently. Procedural competency is threatened by an increasing and diverse amount of procedures that are incorporated in university curricula. As a consequence, less time is available for the learning of each specific procedure. Incorrectly performed procedures in physiotherapy might be ineffective and may result in injuries to patients and physiotherapists. The aim of this review was to synthesise relevant literature systematically to appraise current knowledge relating to assessments for procedural skills in physiotherapy education.
Method: A systematic search strategy was developed to screen five relevant databases (CINAHL, Cochrane Central, SportDISCUS, ERIC and MEDLINE) for eligible studies. The included assessments were evaluated for evidence of their reliability and validity.
Results: The search of electronic databases identified 560 potential records. Seven studies were included into this systematic review. The studies reported eight assessments of procedural skills. Six of the assessments were designed for a specific procedure and two assessments were considered for the evaluation of more than one procedure. Evidence to support the measurement properties of the assessment was not available for all categories.
Discussion: It was not possible to recommend a single assessment of procedural skills in physiotherapy education following this systematic review. There is a need for further development of new assessments to allow valid and reliable assessments of the broad spectrum of physiotherapeutic practice

Abstract

Einleitung: Das Erlernen von prozeduralen Fähigkeiten ist ein wichtiges Element in der Ausbildung von Physiotherapeuten/-innen. Es ist das Ziel von physiotherapeutischen Studiengängen, Graduierte zu befähigen, ausgewählte Prozeduren sicher und effektiv auszuführen. Die prozedurale Kompetenz ist bedroht von wechselnden und einer stetig anwachsenden Anzahl von Prozeduren, die in die Curricula der Studiengänge eingebaut werden. Als Konsequenz ist weniger Zeit vorhanden, um die einzelnen Prozeduren zu erlernen. Falsch durchgeführte Prozeduren können zu Verletzungen von Patienten/-innen und Physiotherapeuten/-innen führen.
Zielsetzung dieser Arbeit war es, relevante Literatur systematisch zu erfassen, um eine Übersicht von Assessments von prozeduralen Fähigkeiten in der physiotherapeutischen Ausbildung zu erstellen.
Methode: Eine systematische Suchstrategie wurde entwickelt, um fünf Datenbanken (CINAHL, Cochrane Central, SportDISCUS, ERIC und MEDLINE) nach relevanten Studien zu durchsuchen. Die eingeschlossenen Assessments wurden bezüglich Reliabilität und Validität bewertet.
Ergebnisse: Die Suche in den elektronischen Datenbanken ergab 560 Treffer. Sieben Studien wurden in diese systematische Übersichtsarbeit eingeschlossen. Die Studien berichteten über acht Assessments für prozedurale Fähigkeiten. Sechs Assessments sind für eine spezifische Prozedur entwickelt worden und zwei Assessments können für unterschiedliche Prozeduren benutzt werden. Evidenz für die Messeigenschaften der eingeschlossenen Messinstrumente war nicht für alle Kategorien verfügbar.
Diskussion: Es ist nicht möglich, ein bestimmtes Messinstrument zur Bewertung von prozeduralen Fähigkeiten zu empfehlen. Es gibt einen Bedarf an Messinstrumenten, die reliabel und valide sind, um das breite Spektrum von prozeduralen Fähigkeiten zu bewerten.

Keywords

Procedural skills – practical skills – systematic review – clinical assessment

Keywords

Prozedurale Fähigkeiten – praktische Fähigkeiten – systematischer Review – klinisches Assessment

INTRODUCTION

It is the aim of physiotherapy degree programmes that graduates are able to execute selected procedures safely and efficiently. Considerable resources are allocated to enable graduates to achieve a high level of procedural competency. Within this review, procedural skills were classified after Kent's definition as: 'a skill involving a series of discrete responses each of which must be performed at the appropriate time in the appropriate sequence' (Kent, 2007, p. 437).

Recent literature highlights that there is no consensus with regard to definitions and classifications of procedural skills. Michels, Evans, and Blok (2012) identified that procedural skills are not exactly defined in the field of health professions education. Frequently, they are categorised under the umbrella term 'clinical skills'. However, there is a lack of standardisation. Simpson et al. (2002) separated the practical procedures from communication skills, clinical skills, and other skills in the Scottish doctor learning outcomes. In contrast, the General Medical Council in the UK does not separate between procedural skills and clinical skills (2004), for example, safety measures are categorised as essential procedural skills in their classification. Lastly, the Royal Australian College of General Practitioners (2011) defined procedural skills as: 'A procedure is a manual intervention that aims to produce a specific outcome during the course of patient care' (The Royal Australian College of General Practitioners, 2011, p . 515).

To avoid ambiguity in this review, procedural skills were characterised with the following features: a) they involve the execution of a procedural task (e.g., a manual or a practical task), b) involvement of technical equipment may be possible but this is not a prerequisite of procedural skills, c) the character of a procedure can be diagnostic, evaluative or interventional and d) procedures can range from simple tasks with few parts to complex sequences involving multiple activities.

As procedures in physiotherapy are highly interactive between patients and therapists, more information than the execution of procedures may be needed to evaluate the procedural skills. For example, communication providing basic information about the procedures between physiotherapist and patient is frequently necessary. Consequently, therapists should be educated to allow them to adapt procedures to a variety of circumstances such as environmental requirements or individual patient needs.

Physiotherapy is a dynamic profession with evolution of new physiotherapeutic roles and skills in many health systems (Higgs, Hunt, Higgs, & Neubauer, 1999), thus requiring the incorporation of new tasks and skills into physiotherapy degree curricula. However, this may result in an increased amount of procedures that are incorporated in university curricula. As a consequence, less time is available for the learning of specific procedures.

Incorrectly performed procedures in physiotherapy might be ineffective and may result in injuries to physiotherapists or to patients. For example, Nyland and Grimmer (2003) reported that low back pain is frequently experienced by undergraduate physiotherapy students and, Glista and co-workers (2014) reported that the students' posture deteriorated during the course of education. In some situations, physiotherapists are required to perform professional procedures in difficult environments with poor working postures which are potentially harmful for the musculoskeletal system (Jackson & Liles, 1994). Therefore, training of procedures should be designed to enable learners to perform procedures without endangering their own personal safety and to understand how to adapt procedures appropriately.

Procedures performed by physiotherapist can also be associated with adverse events for patients. For example, Gorrell, Engel, Brown, and Lystad (2016) reported that mild adverse events occurred in 61 RCTs and major adverse events were seen in 2 RCTs evaluating the spinal manipulative therapy. Therefore, following the initial teaching of procedural skills, physiotherapy educators need valid and reliable assessment tools to evaluate whether procedural competency of students is sufficient for practice.

Assessment of procedural skills has been extensively researched in surgical education (Jelovsek, Kow, & Diwadkar, 2013). Some assessments exist, which can be used for procedures in nursing education (Morris, Gallagher, & Ridgway, 2012). While teaching of procedural skills is a core part of undergraduate physiotherapy education, no review could be identified of assessment tools for procedural skills in physiotherapy education.

One important consideration in the evaluation of procedural skills in physiotherapy is whether an assessment framework exists. Miller (1990) argued that no single assessment would be sufficient to allow the judgement of such complex skills. He presented a four-level framework for assessments in health professions education. The base of this framework is knowledge (the student 'knows'), which can be tested with standardised objective test methods (e.g., multiple choice tests). The second level (competence) provides evidence that students know how to use their knowledge (e.g., vignette assessments). The third level evaluates the performance of students (e.g., students have to show how they perform a specific procedure). Lastly, the question remains whether the learned skills are independently selected and used appropriately in clinical practice. Examples to evaluate the 'action level' are work place based assessments or

portfolios (Chandratilake, Davis, & Ponnamperuma, 2010).

The aim of this review was to identify, examine and synthesise relevant literature to produce a systematic review of assessments for procedural skills in physiotherapy education. Specifically, the objective of this review was to identify existing assessments of procedural skills in physiotherapy education and to evaluate them with regard to their measurement properties.

METHODS

A systematic review was undertaken to address the identified objectives. To increase clarity of reporting, the PRISMA guideline was followed (Liberati et al., 2009).

Criteria for inclusion and exclusion

Inclusion and exclusion criteria are presented in Table 1.

Search methods

Five electronic databases were systematically searched for potential eligible studies. These databases were: Cumulative Index to Nursing and Allied Health Literature (CINAHL), Cochrane Central Register of Controlled Trials (CENTRAL), SPORTDiscus, Educational Resource Information Center (ERIC) and Medline via Pubmed. In addition, the references of all included full text articles were checked for relevant studies. The search string is presented in Table 2. Findings of the three categories Population, Assessment and Outcome were combined with the Boolean operator AND.

All retrieved records were imported into an electronic database and duplicates were removed. In the next step, titles and abstracts of the records were screened with regard to the pre-specified inclusion and exclusion criteria. Lastly, the full texts of the remaining studies were read and studies were included in the systematic review if they met all criteria.

Table 1: In-and exclusion criteria

Category	Criteria
Population	**Studies with physiotherapists or physiotherapy students were included.**
	Studies with health professionals or health professions students were included when they practiced procedures that can be used in physiotherapy (i.e., when medical students were evaluated on their ability to perform a musculoskeletal examination)
	Studies with health professionals or health professions students were excluded when they practiced procedures that cannot be practiced by physiotherapists (such as surgery)
Educational assessments	**The assessment could be either a procedure specific measurement instrument (i.e., the assessment is designed exclusively for one procedure) or a procedure unspecific measurement (i.e., the assessment is designed to measure procedures in physiotherapy education but can be used for more than one procedure)**
	The assessment should measure procedures in reality. Assessments based on virtual reality were excluded.
	The assessment should be feasible in various settings. Therefore, assessments that require expensive equipment were excluded.
	Data must be available for a specific assessment. Studies with summary data of several assessments were excluded (e.g., summary scores of a complete OSCE).
Outcome	**The aim of assessment should be to measure procedural skills. Assessments of similar constructs such as clinical skills or psychomotor skills [defined as '… motor skill, some manipulation of material, or some act which requires a neuromuscular action' Simpson (1966, p. 17)] were included.**
	Assessments that aimed to exclusively evaluate other outcomes such as communication skills or professionalism were excluded.
	When assessments were designed to measure multiple outcomes, it was evaluated whether the focus was based on procedural skills (e.g., more than 50% of the items concentrate on procedural skills). The assessments with focus on procedural skills were included.
Measurement properties	**Studies had to report the measurement properties of an educational assessment (e.g., reliability or validity)**

Table 2: Search strategy

Population	Assessment	Outcome
medical education OR education, medical[Mesh] OR physiotherapy education OR physical therapy education OR health professions education OR healthcare education OR allied health care education	scale OR global rating scale OR GRS OR checklist	practical skill* OR psychomotor skill* OR procedural skill* OR clinical skill*

Data collection and management

Data were extracted in relation to the following information:

- Study details (country, setting and sample)
- Assessment characteristics (name of the assessment, assessment items, assessment aim, assessment duration, assessment criteria, assessors, patients and target procedure)
- Measurement properties (internal consistency, reliability, measurement error, content validity and construct validity)
- Methodological quality of assessments (the Standards for Evaluating the Quality of Assessment Methods in Medical Education (Swing, Clyman, Holmboe, & Williams, 2009)

Analysis

Evidence of reliability and validity of the included assessments was evaluated. Within reliability the internal consistency, the inter- and intra-rater reliability and the measurement error were appraised. Validity was appraised with regard to content validity, criterion validity and construct validity. Despite some discussion about agreed definitions regarding measurement properties, the consensus definitions proposed by Mokkink et al. (2010) were used to ensure consistency in how findings were interpreted.

Assessment of methodological quality of assessments

All included assessments were evaluated with the Standards for Evaluating the Quality of Assessment Methods (SEQAM) (Swing et al., 2009). The SEQAM is an assessment tool for educational assessments specifically designed for health professions education. The SEQAM critically evaluates 6 dimensions: reliability (e.g., reliability indicators are available for all used scores), validity (e.g., selection of content is justified), ease of use (e.g., the tool is easily carried out in daily practice), resources required (e.g., training requirements for assessors do not exceed one hour), ease of interpretation (e.g., individual scores are interpretable) and educational impact (e.g., provides useful results). For each dimension, the studies could be rated as evidence level A, B, C or not rated. For an evidence level of A, all standards of one dimension had to be met. Studies were rated as evidence level B when one standard was not met. When two standards in one dimension were not met, an evidence level of C was specified. Lastly, when three or more standards were not met, an evidence level of not rated (NR) was given. The scoring rules of the SEQAM were adapted from Swing et al. (2009).

RESULTS

The results of this review are presented in three sections. First the results of the search are presented, then the findings of measurement properties of the included assessments are provided. Finally, the methodological quality of the included assessments is considered.

Results of the search

The search of electronic databases identified 560 potential records. Additionally, 10 articles were identified by reference checking. It was possible to delete 6 duplicates. Therefore, titles and abstracts of 564 records were screened. The majority of 454 records were excluded because they did not report an appropriate assessment (n= 387). Fifty records did not report an appropriate outcome and 17 records did not meet the inclusion criteria with regard to the population.
110 full-text articles were then read. It was possible to exclude 103 full-text articles. Most studies (n = 93) were excluded because they were related to a different discipline in medicine (e.g., surgery). Two studies had insufficient data to include them into the systematic review. They evaluated multiple different patient encounters, and therefore, it was not possible to extract data for a single assessment method. Eight studies were not included because they were reviews of primary studies. Finally, seven studies were included into this systematic review. The studies reported six procedure specific measurement instruments (PSMI) and two procedure unspecific measurement instruments (PUMI) (Figure 1).

Included assessments

The included assessments were classified as either procedure specific measurement instruments (i.e., assessments designed for one specific procedure) or procedure unspecific measurement instruments (i.e., generic assessments, which can be used for more than one specific procedure).

Procedure specific measurement instruments

The six PSMIs included in this review are briefly presented below. A detailed critical overview is presented in Table 3. The Assessment of Musculoskeletal Physical Examination Skills Checklist (AMPE) was published by Beran et al. (2012). The AMPE is a 12-15 item checklist and evaluates the ability of health professionals to perform a physical examination of four different clinical scenarios. The scenarios involve an upper extremity, a trauma, a spine and a lower extremity case. The AMPE requires, in addition to an assessor, a trained standardised patient for each of the four scenarios. The authors designed checklists of important procedures, which the students

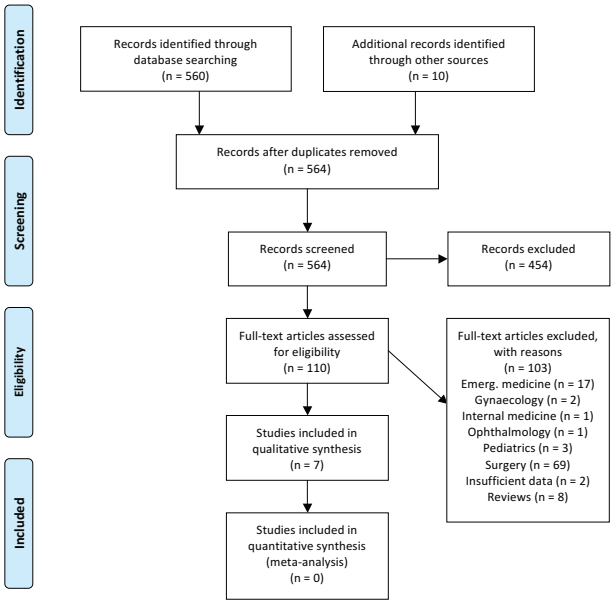

Figure 1: Study flow.

should perform when they encounter a specific simulated patient, such as joint palpation or strength testing.

Herbers, Wessel, El-Bayoumi, Hassan, and St Onge (2003) created the 29-item Pelvic Examination Skills Checklist (PES-C) and the 5-point Pelvic Examination Skill Rating Scale (PES-R). Most of the 29 items on the PES-C are related to the physical performance of a pelvic examination, although some of the items relate to communication skills (e.g., item 21: Tells patient to state if pain too great). The PES-R is a five-point global rating scale that enables the evaluator to rate the overall performance of the pelvic examination. Both assessments were validated with gynaecologic teaching associates who fulfilled a dual role as subjects for the pelvic examination and evaluators of the learner's performance within the study of Herbers and colleagues.

The Physical Examination Skills Checklist (PhyES) was published by Ladyshewsky, Baker, Jones, and Nelson (2000) and aims to evaluate a musculoskeletal physical examination of a patient with a rotator cuff problem. The PhyES is scored on a three-point system and uses carefully coached persons to portray specific patients. Performance was scored using a checklist which included important features of the physical examination (e.g., evaluation of shoulder girdle stability).

Swift and colleagues (2013) designed the mOSCE-Station 3 checklist (mO-S3). The mO-S3 evaluates the ability of physiotherapy students to perform two specific shoulder assessment tests. Learners have to choose two tests to confirm their hypothesis with regard to a scenario with a patient suffering from shoulder pain. The mO-S3 consists of five dichotomous items and one ordinal item. In order to administer the mO-S3, standardised patients and specialised clinical instructors are necessary. The following tasks were evaluated in the OSCE: i) think station, ii) explanation of the primary hypothesis to a patient, iii) performing two specific tests to confirm the hypothesis, iv) performing the best day 1 hands-on intervention, v) reassessment, vi) performing the best day 1 exercise intervention and vii) performing a specific technique and explanation of the selected technique.

The 138 item checklist head- to-toe physical examination checklist (HTTPE) (Yudkowsky et al., 2004) evaluates the ability of an 'assessor' to perform a complete physical screening examination of the whole body and all 138 items are scored on a trichotomous scoring system. To administer the HTTPE, trained standardised patient instructors are required. The patient instructors serve as patients and mark the 'assessors' performance.

Procedure unspecific measurement instruments

The Osteopathic Manipulative Treatment assessment tool (OMT) (Boulet, Gimpel, Dowling, & Finley, 2004) aims to measure the ability to perform a manipulative treatment and consists of 15 items scored on a trichotomous scale. It can be used for different manipulative treatment techniques and for different body regions and therefore is procedure unspecific. For example, Boulet et al. (2004) used the OMT to evaluate various procedures related to the treatment of low back pain, frozen shoulder or asthma. Standardised patients are a prerequisite to use the OMT as an assessment tool.

The Global Procedural Skills Evaluation Form (GPSE) was originally presented in the field of family medicine (Nothnagle, Reis, Goldman, & Diemers, 2010). However, its generalised design as a rating scale for procedural skills affords its utility for the assessment of procedural skills in physiotherapy as well. The GPSE provides feedback based on direct observation of a procedure. The scoring system is based on a 4-point scale and quantifies the amount of guidance that was needed to perform a procedure. No standardised patients are required when the GPSE is applied. Furthermore, student's self-assessment is included in the GPSE score.

Findings

Within this section, the evidence of measurement properties of the included assessments are presented. The consensus definitions proposed by Mokkink et al. (2010) were used to appraise the measurement properties.

Reliability

Reliability of the assessments was appraised with regard to their internal consistency, inter-rater reliability, intra-rater reliability and measurement error.

Table 3: Characteristics of included studies and assessments.

Study	Country	Setting	Sample	Assessed procedure	Scale and items	Duration	Patients	Assessors	Purpose
Beran et al. (2012) AMPE	USA	Orthopaedic department	24 orthopaedic residents	PSMI: Musculoskeletal physical examination; Inspection, palpation, joint range of motion, strength testing and any special tests pertinent to the clinical scenario	Four 12-15 items checklists for clinical scenarios (upper extremity, lower extremity, trauma and spine) on dichotomous scales (yes or no).	10 Minutes	Standardised patients are required (120 minutes training)	Pool of experienced raters	High stakes purpose
Boulet et al. (2004) OMT	USA	Osteopathic college	121 osteopathic students (4th year)	PUMI: Osteopathic manipulative treatment of three clinical cases (low back pain, frozen shoulder and asthmatic with cough)	OMT (Osteopathic Manipulative Treatment) assessment tool with 15 items; Every item is scored on a 0 to 2 scale (0 = done incorrectly or not done, 1 = not performed optimally and 2 = done proficiently)	13 minutes	Standardised patients with 8 hours of formal training	16 osteopathic physicians (5 hours of formal training)	High stakes examination (OSCE)
Herbers et al. (2003) PES-C	USA	University Medical Centre	72 internal medicine residents	PSMI: Pelvic examination	29 item dichotomous checklist (yes = when the behaviour was observed; no = when the behaviour was not observed); Includes some items about communication skills	Not specified	Gynaecologic teaching trainer required; 1 trainer was being examined and the second trainer rated the student's skills.	Gynaecologic teaching trainer required	Not specified
Herbers et al. (2003) PES-R	USA	University Medical Centre	72 internal medicine residents	PUMI: Pelvic examination	Global rating scale evaluating the overall performance of the pelvic examination (five-point ordinal scale between 1 = inadequate and 5 = excellent)	Not specified	Gynaecologic teaching trainer required; 1 trainer was being examined and the second trainer rated the student's skills	Gynaecologic teaching trainer required	Not specified
Ladyshewsky et al. (2000) PhyES	Australia	Physiotherapy department	12 under-graduate physiotherapy students 4 physio-therapists (at least 2 years of experience)	PSMI: Musculoskeletal physical examination of a patient with a rotator cuff problem	Physical examination checklist (3-point scale: 0 = not done, 1 = done poorly or incompletely and 2 = done well), number of items not available	Mean 30 minutes (range: 20 - 46 minutes)	Standardised patients are required	Assessors with 30 hours of training	High stakes examination (OSCE)

Continued *Table 3: Characteristics of included studies and assessments.*

Study	Country	Setting	Sample	Assessed procedure	Scale and items	Duration	Patients	Assessors	Purpose
Nothnagle et al. (2010) GPSE	USA	Family medicine department	5 faculty members and 5 students (semi structured interviews); Focus groups: 7 experienced family medicine educators, 5 residents and 5 faculty members	PUMI: Eligible for all procedures in family medicine	Global Procedural Skills Evaluation Form, 4-point scale, amount of assistance is documented ranging from significant guidance is provided to performed independently; communication skills etc. are included; Student's self-assessment is included; Difficulty of the procedures is rated as well	Not specified	Not required	Not specified	Low stakes examination (formative feedback)
Swift et al. (2013)* mO-S3	USA	Physiotherapy department	12 undergraduate 1st year physiotherapy students	PSMI: Examination skills in musculoskeletal physiotherapy (shoulder tests)	Checklist for a musculoskeletal OSCE station; 6 items checklist (5 dichotomous items and 1 ordinal item)	6 minutes	Simulation patients with 2 hours of supervised training and 1 week of independent training	Clinical instructors (2 - 20 years of experience)	Low stakes examination (mid-term)
Yudkowsky et al. (2004) HTTPE	USA	University Medical Centre	369 medical students (2nd year)	PSMI: Head to toe physical examination	138 item checklist; three-point scale (0 = incorrect, 1 = correct after prompt, 2 = correct without prompting); Test duration: 2 h; high stakes summative assessment or low stakes formative assessment	2 hours (45 minutes unprompted exam, remaining 1:45 hours are used for scoring, feedback, and teaching)	Trained patient instructors with 25 hours of training	Trained patient instructors with 25 hours of training	High stakes summative assessment and low stakes formative assessment

AMPE: Assessment of Musculoskeletal Physical Examination Skills Checklist; GPSE: Global Procedural Skills Evaluation Form; HTTPE: Head to Toe Physical Examination; mO-S3: mOSCE-Station 3 checklists; PhyES: Physical Examination Skills Checklist; PES-C: Pelvic Examination Skills Checklist; PES-R: Pelvic Examination Skill Rating Scale; PSMI: Procedure Specific Measurement Instrument; PUMI: Procedure Unspecific Measurement Instrument

** Swift et al. (2013): It was only possible to use data from a small pilot study. The follow up study evaluated a 6 station OSCE. Single values for a specific scale were not available.*

Two studies were included that reported the internal consistency of two different assessments. Swift et al. (2013) reported an internal consistency between $\alpha = 0.31$ (video examiner) and $\alpha = 0.55$ (onsite examiner) for the mO-S3. They calculated the internal consistency of a 6 station OSCE. The statistical method used to calculate the internal consistency was Cronbach's alpha. Boulet et al. (2004) reported an internal consistency for the OMT between 0.83 (Case 1: low back pain) and 0.97 (Case 3: asthma). All internal consistency estimates are presented in Figure 2.

Six studies were included that reported the inter-rater reliability of six assessments. Beran et al. (2012) evaluated four different procedures using the AMPE. Inter-rater reliability ranged between 0.27 (95%CI: 0 to 0.56) for the physical examination of trauma patients to 0.77 (95% CI: 0.46 to 0.9) for a physical examination of the knee.

Herbers et al. (2003) investigated the interrater reliability of students performing a specific pelvic examination with no deviations from the protocol allowed and reported kappa coefficient of $\kappa = 0.54$ for the PES-C (pelvic examination).

Ladyshewsky et al. (2000) investigated the interrater reliability for the assessment of a musculoskeletal shoulder examination using the PhyES. A kappa coefficient of $\kappa = 0.79$ was reported.

Swift et al. (2013) published an ICC of 0.77 for the interrater reliability of the mO-S3 based on the clinical competency of doctoral physical therapy students halfway through their education in musculoskeletal physiotherapy.

An interrater reliability of ICC = 0.95 for students scored on all 138 items on the head to toe examination (HTTPE) was reported by Yudkowsky et al. (2004). Lastly, Boulet et al. (2004) reported a correlation coefficient of r = 0.83 (range r = 0.06 - r = 0.93) for the interrater reliability of the OMT. The authors reported that the average difference between two raters was 2.4 points on a 0 to 30 points scale. All interrater reliability estimates are presented in Figure 3.

Intra-rater reliability was available for only one assessment. Ladyshewsky et al. (2000) published an intra-rater reliability of $\kappa = 0.63$ for the PhyES.

None of the studies included in this review evaluated the measurement error of their included assessments.

Validity

Validity of the included assessments was evaluated with regard to their content validity, criterion validity and construct validity.

Evidence for content validity was found for four assessments AMPE, PhyES, GPSE and mO-S3 (Beran et al., 2012; Ladyshewsky et al., 2000; Nothnagle et al., 2010; Swift et al., 2013). For each assessment, the authors

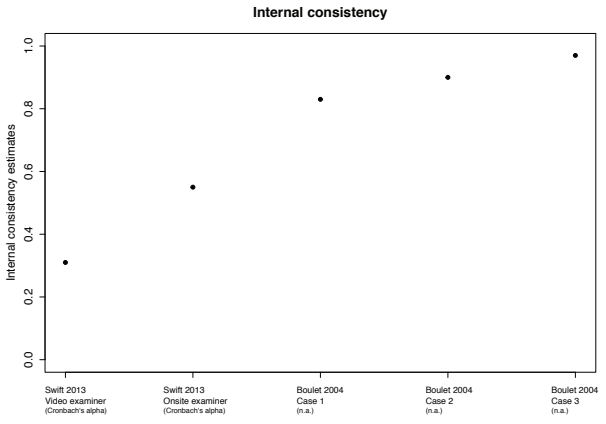

Figure 2: Internal consistency estimates.
Nb. The statistical method from Boulet et al. (2004) was not available.

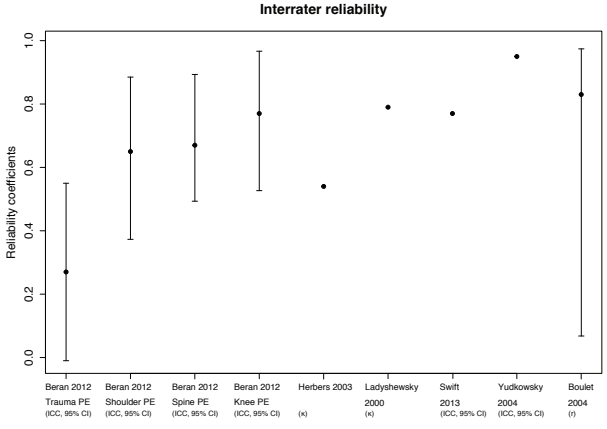

Figure 3: Interrater reliability estimates.

provided information about how their assessments were designed. All four studies used expert panels to judge the comprehensiveness and relevance of the assessment items. The size of the expert panels ranged between an unspecified number of panel members for the AMPE and mO-S3 (Beran et al., 2012; Swift et al., 2013) to 17 participants for the GPSE (Nothnagle et al., 2010). Additionally, two studies involved learners in the process of designing the assessment PhyES and GPSE (Ladyshewsky et al., 2000; Nothnagle et al., 2010) with Nothnagle et al. (2010) generating content for the GPSE through three focus groups. None of the studies within this review reported the criterion validity of their assessments. Therefore, the utility of using the assessments to predict future performance or as compared to another measure is not known.

Data regarding the construct validity was available for five assessments AMPE, OMT, PES-C, PES-R, PhyES (Beran et al., 2012; Boulet et al., 2004; Herbers et al., 2003; Ladyshewsky et al., 2000). Three studies tested the

hypotheses whether their assessments could discriminate performance between individuals with more experience or less experience. Beran et al. (2012) reported that years of training had no significant influence on the total score of the AMPE. Ladyshewsky and colleagues found in their study that licenced physiotherapists performed significantly better on the PhyES than fourth year undergraduate students. Lastly, Herbers et al. (2003) presented the evidence that learners in a training group scored significantly higher than learners without a specific training ($p< 0.001$) on the PES-C. Two studies reported correlations between the included assessments and the other established assessments as evidence for construct validity. Herbers et al. (2003) reported an agreement of $K = 0.66$ between their checklist for a pelvic examination (PES-C) and a global rating scale for this procedure (PES-R). Boulet et al. (2004) reported that the OMT instrument correlated with biomedical knowledge indicators ($r = 0.47$) and global patient assessment ($r = 0.46$).

Methodological quality of assessments

Methodological quality of the included assessments was low to moderate. Methodological quality was appraised with 20 standards of the SEQAM. The assessment that was appraised as fulfilling the most standards was the AMPE. Ten of the 20 standards were appraised as fulfilled. The mO-S3 was evaluated as fulfilling the least standards (7 standards were classified as satisfied). All standards are presented in Table 4.

DISCUSSION

The discussion is divided into the following sections: 1) summary of main results, 2) methodological quality of the assessments, 3) potential biases in the review process, and 4) agreements and disagreements with other studies.

Summary of main results

This systematic review synthesised relevant literature relating to the current knowledge of assessments for procedural skills in physiotherapy education. Following a systematic search, eight assessments for procedural skills were identified that can be used in physiotherapy education. Six of the assessments were designed for a specific procedure and were validated for diagnostic or evaluative procedures. Two assessments (GPSE and OMT) were considered useful for the evaluation of more than one procedure and can be used to evaluate procedural competence of therapeutic interventions.

The GPSE was classified as representing the highest level of Miller's framework of assessments (Miller, 1990) and can be used as a workplace based assessment, which is the 'Does' level in Miller's pyramid. All the remaining assessments were classified as representing the 'Shows how' level, because they were all based in a simulated environment and no direct evidence was available to evaluate whether the behaviour of the learners actually changed.

In terms of internal consistency, the best performing assessment, (OMT), had a value above 0.70, while the other assessment reporting internal consistency (the mO-S3) had lower estimates. These lower values of the mO-S3 might be explained by the method to calculate internal consistency which was used by Swift et al. (2013). They calculated internal consistency with regard to a 6 station OSCE, with stations designed to measure competence in musculoskeletal physiotherapy. However, the content of the stations varied to some extent. This conflicts with the stance of Cortina (1993) who stated that when internal consistency is measured, the set of test items should form a reflective model, that is, 'all items are a manifestation of the underlying construct' (Mokkink et al., 2009, p. 24). It could be argued that the stations and test items of the OSCE devised by Swift et al. (2013) did not measure the same construct (e.g., diagnostic, interventional or communication competence) or that they measured different aspects of one construct. This could explain the lower internal consistency estimates of the mO-S3.

Six of the included assessments reported inter-rater reliability. The highest estimate was reported for the HTTPE (ICC: 0.95). The AMPE and the PES-C were evaluated as having moderate to low inter-rater reliability because estimates were below 0.70. There are a number of methodological issues that may have affected the reliability. For the PES-C, Herbers et al. (2003) calculated their reliability scores based on a subset of their items (i.e., only data of 7 of the 29 items of the PES-C were used). Additionally, the study used audiotapes to calculate the reliability between the two raters. With regard to a checklist that aims to evaluate procedural skills, important issues may have been missed, which can only be detected visually. Therefore, only such items as: 'Asks if patient wants mirror to watch examination' were evaluated with regard to their reliability. In relation to the AMPE, three out of the total of four different assessments scored around or above the 0.7-margin. Only the AMPE assessment of a physical examination of trauma patients scored considerably lower (ICC = 0.27). Beran et al. (2012) reported that considerable disagreement was present between the raters. One rater scored consistently higher than the two other raters. In an attempt to improve the reliability, the scores of three raters were averaged and compared with an external rating. This method resulted in increased interrater reliability scores (ICC = 0.51).

Table 4: Methodological quality of included assessments.

Standards for evaluating the quality of assessment methods	Beran 2012 (AMPE)	Boulet 2004 (OMT)	Herbers 2003 (PES-C&R)	Ladyshewsky 2000 (PhyES)	Nothnagle 2010 (GPSE)	Swift 2013 (mO-S3)	Yudkowsky 2004 (HTTPE)
Reliability							
1. Reliability indicators	☺	☺	☹	☺	☹	☺	☺
2. Inter- and Intra-rater reliability	☹	☹	☹	☺	☹	☹	☹
3. High-stakes decisions	☹	☹	☹	☹	☹	-	☹
Level of evidence (A, B, C or NR)	C	C	NR	B	NR	C	C
Validity							
1. Interpretation of results	☺	☺	☺	☺	☺	☹	☺
2. Selection of content	☺	☹	😐	☺	☺	☺	😐
3. Unintended consequences	☺	☺	☺	☺	😐	☺	☺
4. Agreement between a single expert and consensus ratings	☺	☹	☹	☹	☹	☹	☹
5. Subjective judgment	☹	☹	☹	☹	☹	☹	☹
Level of evidence (A, B, C or NR)	B	NR	NR	C	NR	NR	NR
Ease of use							
1. Daily practice	☹	☹	😐	☹	☺	☹	☺
2. Special set up	☺	☺	☺	☺	☺	☺	☺
3. Duration	☺	☺	😐	☹	😐	☺	☹
Level of evidence (A, B, C or NR)	B	B	C	C	B	B	B
Resources required							
1. Additional resources	☺	☺	☺	☺	☺	☺	☺
2. Training requirements	☹	☹	☹	☹	😐	☹	☹
3. Additional persons	☹	☹	☹	☹	☺	☹	☹
Level of evidence (A, B, C or NR)	C	C	C	C	B	C	C
Ease of interpretation							
1. Interpretation of individual scores	☺	☺	☺	😐	☺	☹	☺
2. Normative data	☺	☹	☺	☹	☹	☹	☹
3. Individual to group performance.	☹	☹	☹	☹	☹	☹	☹
Level of evidence (A, B, C or NR)	B	C	B	NR	C	NR	C
Educational impact							
1. Positively affect individual learners	☹	☹	☺	☹	☹	☹	☺
2. Positively affect programme curriculum	☹	☹	☹	☹	☹	☹	☹
3. Provide useful results	😐	😐	😐	😐	☺	😐	☺
Level of evidence (A, B, C or NR)	NR	NR	C	NR	C	NR	B

A level of evidence A was assigned when all standards in one dimension were met. A level of B was assigned when one standard was not met. A level of C was appraised when two standards were not met and NR was assigned when more than two standards were not met. ☹: Standard not met; ☺: Standard met; 😐: Unclear; -: Standard not applicable

Only the PhyES evaluated the intra-rater reliability, reporting a moderate agreement ($\kappa = 0.63$). These findings should be interpreted with caution due to the very small sample (six encounters over two occasions during a two-week period).

When a new assessment is developed, users require reassurance that the instrument is comprehensive and relevant. This might be assured by using experts to comment on or generate the content of the assessment (Mokkink et al., 2009). Furthermore, the proposed assessment should match the target population with regard to focus and detail, and one way of assuring this is to recruit potential participants and discuss the assessment with them. However, only the PhyES (Ladyshewsky et al., 2000) and the GPSE (Nothnagle et al., 2010) included students into the design of the assessments. Nothnagle et al. (2010) also used a more robust development process, including focus groups, to construct their assessment (GPSE), which may make it more likely that this assessment is comprehensive and consists of relevant items.

Evidence of construct validity was found for four assessments (PES-C, PES-R, PhyES and OMT). It has been established that learners should improve execution of a procedure in response to the level of experience and increased amounts of practice (Brydges, Carnahan, Backstein, & Dubrowski, 2007). Specifically, the PES-C and the PhyES were able to differentiate between learners with different levels of experience; however, this was not established for the AMPE.

Methodological quality of assessments

Methodological quality of assessments was evaluated with the SEQAM, which is based on the utility index of Van Der Vleuten (1996). The author argued that the appraisal of assessment methods in health professions education should consider more than traditional measurement properties (i.e., reliability and validity). Within his utility index he stressed the importance of the acceptability, the educational impact and the cost effectiveness of an assessment. The educators should take this information into account when context specific decisions about assessments are made (Van Der Vleuten & Schuwirth, 2005). Similarly, the SEQAM critically evaluates six dimensions: reliability, validity, ease of use, resources required, ease of interpretation and educational impact.

Overall, the methodological quality of the included assessments was low to moderate (fulfilling between 6 and 10 standards). No assessment was appraised as having no risk of bias. No study fulfilled all educational standards of the SEQAM. The assessment that was appraised as fulfilling the most standards was the AMPE with 10 of the 20 standards fulfilled. The mO-S3 was evaluated as fulfilling the least standards (6/20). The remaining assessments ranged between seven to nine standards fulfilled. One reason for this moderate quality of evidence was that it was derived from only a single study for each assessment. Therefore, it was not possible to complete some standards (e.g., the item 'positively affects programme curriculum' can only be awarded if at least two studies present the evidence).

A discrepancy existed between the assessment and the standard 'training requirements'. The standard sets the benchmark for training time to one hour, in order to reduce the required resources. In contrast, most of the researchers spent considerably more time in the training of faculty members and standardised patients, with Ladyshewsky et al. (2000) spending up to 30 hours in the training of their assessors. This is not viable in an educational programme, and therefore, finding a reasonable balance between those extremes will be a challenge for further work.

Within the 'non-traditional' categories of measurement properties (e.g., non-psychometric properties), it was noted that five assessments were classified as 'relatively easy to use' because they required little specialist set up and time to evaluate (Beran et al., 2012; Boulet et al., 2004; Nothnagle et al., 2010; Swift et al., 2013; Yudkowsky et al., 2004). However, only the GPSE was appraised as also requiring few resources (Nothnagle et al., 2010). This could be important for educators when they need assessments in their daily practice, which are easy to set up and use.

Potential biases in the review process

Only one study for each assessment was identified; hence, limiting generalisability and rendering it impossible to perform a meta-analysis. Findings have therefore been presented narratively. Furthermore, sample size may affect findings, only three studies evaluated their assessments with considerable sample sizes. Boulet et al. (2004), Herbers et al. (2003), and Yudkowsky et al. (2004) used at least 70 participants in their studies. The remaining studies recruited considerably fewer (< 25) participants, which again may limit generalisability and may have caused imprecision of the effect estimates.

A cut off value of 0.7 was used for the measurement properties of internal consistency and inter-rater reliability and intra-rater reliability (Terwee et al., 2007). While other authors use different cut off values (e.g., 0.85 cut off) (Weiner and Stewart (1998), the more moderate interpretation was selected as 0.85 may be too high to be useful in practical settings (Streiner, Norman, & Cairney, 2014). An acceptable reliability standard should be chosen with regard to a specific situation. In high stake examinations (i.e., tests with serious consequences for the tester in situations such as education or certification (Sackett, Schmitt, Ellingson, & Kabin, 2001)), higher reliability is required as compared to a low stakes examinations (i.e., tests without serious consequences for the learner).

A further potential bias in this review is that the SEQAM grading of the methodological quality of assessment was modified. Swing et al. (2009) originally suggested an overall recommendation (i.e., class of evidence) based on the evidence levels provided for each dimension. We decided against the use of an overall score because firstly, in our view, scores should only be combined when they are unidimensional (i.e., the same attribute of the object 'methodological quality' should be measured with different sub-categories) and evidence for unidimensionality was not available for SEQAM; secondly, the use of summary scores might lead to biased estimates in systematic reviews and meta-analysis (da Costa, Hilfiker, & Egger, 2013; Juni, Altman, & Egger, 2001). Therefore, we decided to omit the overall recommendations and present relevant methodological aspects individually.

Agreements and disagreements with other studies or reviews

Four recent systematic reviews were identified that reported the assessment of procedural skills in health professions education (Bould, Crabtree, & Naik, 2009; Jelovsek et al., 2013; McKinley et al., 2008; Morris et al., 2012).

In general, these reviews focussed on medical education and few assessments relevant for use by allied health professions were identified. For example, of the assessments evaluated in this review, only the OMT scale was identified by McKinley and colleagues. The remaining assessments were not discussed in other reviews. Existing reviews do however agree that there is a lack of assessments for procedural skills in allied health profession. In contrast, a considerably greater number of assessments are available for use in medical education: McKinley et al. (2008) included 85 different scales in their review of assessments used in medical education. Our findings were similar to those of Jelovsek et al. (2013), who found that there was limited reporting of measurement properties. Bould et al. (2009) suggested that procedure unspecific assessments tended to miss errors in safety issues. We were not able to comment as only two procedure unspecific assessments were included in this review, and this is therefore an area where uncertainty remains and further work is required.

CONCLUSION AND IMPLICATIONS

Following this systematic review, it was not possible to recommend a single assessment of procedural skills in physiotherapy education; all the assessments we identified have elements of strengths and weaknesses. Therefore, evaluators should use existing tools carefully when evaluating the procedural performance of physiotherapy students. Most assessments we identified were developed for use within the speciality of musculoskeletal physiotherapy and these could be integrated into educational practice. There is, however, a need to develop new assessments to allow valid and reliable assessments of the broader spectrum of physiotherapeutic practice in other specialities (e.g., neurological practice and respiratory practice). When assessments are selected or developed, faculty members should carefully consider issues such as the usefulness and possible interpretation of the findings as well as the more well established focus on measurement properties such as validity and reliability. This may help prevent neglect of issues of importance to relevant stakeholders. Future studies aiming to design new assessments should involve all stakeholders in the design of the content, use and scoring of the assessment. Furthermore, the construct(s) to be measured should be clearly defined.

References

Beran, M. C., Awan, H., Rowley, D., Samora, J. B., Griesser, M. J., & Bishop, J. Y. (2012). Assessment of musculoskeletal physical examination skills and attitudes of orthopaedic residents. The Journal of Bone & Joint Surgery, 94(6), e36 31-38.

Bould, M. D., Crabtree, N. A., & Naik, V. N. (2009). Assessment of procedural skills in anaesthesia. Br J Anaesth, 103(4), 472-483.

Boulet, J. R., Gimpel, J. R., Dowling, D. J., & Finley, M. (2004). Assessing the ability of medical students to perform osteopathic manipulative treatment techniques. JAOA: Journal of the American Osteopathic Association, 104(5), 203-211.

Brydges, R., Carnahan, H., Backstein, D., & Dubrowski, A. (2007). Application of motor learning principles to complex surgical tasks: searching for the optimal practice schedule. J Mot Behav, 39(1), 40-48. doi:10.3200/jmbr.39.1.40-48

Chandratilake, M., Davis, M., & Ponnamperuma, G. (2010). Evaluating and designing assessments for medical education: the utility formula. Int J Med Educ, 1(1), 1-17.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of applied psychology, 78(1), 98.

da Costa, B. R., Hilfiker, R., & Egger, M. (2013). PEDro's bias: summary quality scores should not be used in meta-analysis. Journal of clinical epidemiology, 66(1), 75.

General Medical Council. (2004). The New Doctor: Guidance on PRHO training. London: GMC.

Glista, J., Pop, T., Weres, A., Czenczek-Lewandowska, E., Podgórska-Bednarz, J., Rykała, J., . . . Rusek, W. (2014). Change in anthropometric parameters of the posture of students of physiotherapy after three years of professional training. BioMed research international, 2014.

Gorrell, L. M., Engel, R. M., Brown, B., & Lystad, R. P. (2016). The reporting of adverse events following spinal manipulation in randomized clinical trials—a systematic review. The Spine Journal.

Herbers, J. E., Jr., Wessel, L., El-Bayoumi, J., Hassan, S. N., & St Onge, J. E. (2003). Pelvic examination training for interns: a randomized controlled trial. Acad Med, 78(11), 1164-1169.

Higgs, J., Hunt, A., Higgs, C., & Neubauer, D. (1999). Physiotherapy education in the changing international healthcare and educational contexts. Advances in Physiotherapy, 1(1), 17-26.

Jackson, J., & Liles, C. (1994). Working postures and physiotherapy students. Physiotherapy, 80(7), 432-436.

Jelovsek, J. E., Kow, N., & Diwadkar, G. B. (2013). Tools for the direct observation and assessment of psychomotor skills in medical trainees: a systematic review. Med Educ, 47(7), 650-673.

Juni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. British Medical Journal, 323(7303), 42.

Kent, M. (2007). The Oxford Dictionary of Sports Science & Medicine (3 ed.). Oxford: Oxford University Press.

Ladyshewsky, R., Baker, R., Jones, M., & Nelson, L. (2000). Evaluating clinical performance in physical therapy with simulated patients. Journal of Physical Therapy Education, 14(1), 31.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gotzsche, P. C., Ioannidis, J. P., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Medicine, 6(7), e1000100. doi:10.1371/journal.pmed.1000100.

McKinley, R. K., Strand, J., Ward, L., Gray, T., Alun–Jones, T., & Miller, H. (2008). Checklists for assessment and certification of clinical procedural skills omit essential competencies: a systematic review. Med Educ, 42(4), 338-349.

Michels, M. E., Evans, D. E., & Blok, G. A. (2012). What is a clinical skill? Searching for order in chaos through a modified Delphi process. Med Teach, 34(8), e573-e581.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. Academic medicine, 65(9), S63-67.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . De Vet, H. C. (2009). The COSMIN checklist manual. Amsterdam: VU University Medical Centre.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of clinical epidemiology, 63(7), 737-745.

Morris, M. C., Gallagher, T. K., & Ridgway, P. F. (2012). Tools used to assess medical students competence in procedural skills at the end of a primary medical degree: a systematic review. Med Educ Online, 17.

Nothnagle, M., Reis, S., Goldman, R., & Diemers, A. (2010). Development of the GPSE: a tool to improve feedback on procedural skills in residency. Fam Med, 42(7), 507-513.

Nyland, L. J., & Grimmer, K. A. (2003). Is undergraduate physiotherapy study a risk factor for low back pain? A prevalence study of LBP in physiotherapy students. BMC musculoskeletal disorders, 4(1), 22.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. American Psychologist, 56(4), 302.

Simpson, J., Furnace, J., Crosby, J., Cumming, A., Evans, P., David, M. F. B., . . . McLachlan, J. (2002). The Scottish doctor--learning outcomes for the medical undergraduate in Scotland: a foundation for competent and reflective practitioners. Med Teach, 24(2), 136-143.

Simpson, E. J. (1966). The Classification of Educational Objectives, Psychomotor Domain.

Streiner, D. L., Norman, G. R., & Cairney, J. (2014). Health measurement scales: a practical guide to their development and use: Oxford university press.

Swift, M., Spake, E., & Gajewski, B. J. (2013). The Reliability of a Musculoskeletal Objective Structured Clinical Examination in a Professional Physical Therapist Program. Journal of Physical Therapy Education, 27(2), 41.

Swing, S. R., Clyman, S. G., Holmboe, E. S., & Williams, R. G. (2009). Advancing resident assessment in graduate medical education. Journal of graduate medical education, 1(2), 278-286.

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., . . . de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. Journal of clinical epidemiology, 60(1), 34-42.

The Royal Australian College of General Practitioners. (2011). Procedural Skills.

Van Der Vleuten, C. P. (1996). The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education, 1(1), 41-67.

Van Der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: from methods to programmes. Med Educ, 39(3), 309-317.

Weiner, E. A., & Stewart, B. J. (1998). Assessing individuals: Brooks/Cole Publishing Company.

Yudkowsky, R., Downing, S., Klamen, D., Valaski, M., Eulenberg, B., & Popa, M. (2004). Assessing the head-to-toe physical examination skills of medical students. Med Teach, 26(5), 415-419. doi:10.1080/01421590410001696452