

Physiotherapy Theory and Practice

An International Journal of Physical Therapy

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/iptp20>

Development and validation of a quality appraisal tool for validity studies (QAVALS)

Shweta Gore, Allon Goldberg, Min H. Huang, Michael Shoemaker & Jennifer Blackwood

To cite this article: Shweta Gore, Allon Goldberg, Min H. Huang, Michael Shoemaker & Jennifer Blackwood (2021) Development and validation of a quality appraisal tool for validity studies (QAVALS), *Physiotherapy Theory and Practice*, 37:5, 646-654, DOI: [10.1080/09593985.2019.1636435](https://doi.org/10.1080/09593985.2019.1636435)

To link to this article: <https://doi.org/10.1080/09593985.2019.1636435>



Published online: 27 Jun 2019.



Submit your article to this journal [↗](#)



Article views: 246



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)



Development and validation of a quality appraisal tool for validity studies (QAVALS)

Shweta Gore ^a, Allon Goldberg^b, Min H. Huang^b, Michael Shoemaker^c, and Jennifer Blackwood^b

^aDepartment of Physical Therapy, School of Health and Rehabilitation Sciences, MGH Institute of Health Professions, Boston, MA, USA;

^bDepartment of Physical Therapy, School of Health Professions and Studies, University of Michigan-Flint, Flint, MI, USA; ^cDepartment of Physical Therapy, College of Health Professions, Grand Valley State University, 1 Campus Dr, Allendale, MI, USA

ABSTRACT

Background: Appraisal of methodological quality of included studies is an important component of conducting systematic reviews. Although several quality appraisal tools are available for intervention studies, fewer tools are available for non-randomized designs, especially for studies of measurement properties.

Objectives: The purpose of this study was to develop a quality appraisal tool specific to validity studies (QAVALS) and to examine its reliability and validity.

Methods: Following identification of key concepts, an initial list of 34 possible items was developed. Content experts rated each item as either ‘essential’, ‘useful but not essential’, and ‘not necessary’. The content validity ratio (CVR) and content validity index (CVI) were calculated to establish content validity following two rounds of review. Inter-rater and test–retest reliability were assessed by two external reviewers using weighted kappa coefficients.

Results: Items below a CVR of 0.50 were eliminated resulting in the modified version with 27 items. Following the second round, the final tool with 24 items was developed. The content validity index of QAVALS was 0.90. QAVALS demonstrated excellent test–retest reliability ($k = 0.80\text{--}0.84$, 95% CI = 0.76–0.90) and good overall inter-rater reliability ($k = 0.70$, 95% CI = 0.61–0.79).

Limitations: Individual item reliability was low for four items. Further research is warranted to examine reliability using larger number of studies and raters with different experience levels.

Conclusion: QAVALS is the first quality appraisal tool specifically designed to address common types of validity. The QAVALS demonstrates strong content validity, good overall inter-rater and excellent test–retest reliability.

ARTICLE HISTORY

Received 15 August 2018

Revised 27 March 2019

Accepted 25 May 2019

KEYWORDS

COPD; chronic obstructive pulmonary disease; questionnaires; validation; physical activity

Introduction

One of the most important aspects of conducting a systematic review is assessing the quality and risk of bias of included studies. Bias within the studies included in a systematic review is detrimental to the overall quality of evidence produced by the review (Kim et al., 2013). Since quality of included studies can vary greatly in terms of internal validity, external validity and the quality of reporting, quality appraisal of studies forms an integral component of a systematic review (Crowe and Sheppard, 2011; Drucker, Fleming, and Chan, 2016; Gopalakrishnan and Ganeshkumar, 2013; Institute of Medicine (US) Committee on Standards for Systematic Reviews of Comparative Effectiveness Research, 2011; Jarde, Losilla, Vives, and Rodrigo (2013); Whiting et al., 2003). Quality appraisal is even more important for non-randomized designs as these are more susceptible to bias as compared to randomized

controlled trials (Jarde, Losilla, Vives, and Rodrigo, 2013). Valid and reliable quality appraisal tools provide a high level of rigor when evaluating the available literature (Jarde, Losilla, Vives, and Rodrigo, 2013). While several quality appraisal checklists are available for assessment of randomized designs (Crowe and Sheppard, 2011; Downs and Black, 1998), fewer tools are available to assess the quality of non-randomized designs, specifically for studies of measurement properties (Lucas, Macaskill, Irwig, and Bogduk, 2010).

Validity studies are types of non-randomized designs that examine the ability to make inferences from measurements and are of several types including face, content, construct, criterion validity, and diagnostic accuracy (Portney and Watkins, 2009). Specific types of validity studies have their own unique designs which follow different methods and analyses.

Currently available validated quality appraisal tools for studies of measurement properties include the quality

appraisal of reliability studies (QAREL) and the quality assessment of diagnostic accuracy studies (QUADAS and QUADAS-2) (Lucas, Macaskill, Irwig, and Bogduk, 2010; Whiting et al., 2003, 2011). The QAREL is specific to studies of reliability (Lucas, Macaskill, Irwig, and Bogduk, 2010). The QUADAS and QUADAS-2, are specific to diagnostic accuracy and therefore do not address all the aspects of validity (Whiting et al., 2003, 2011). The consensus-based standards for the selection of health measurement instruments (COSMIN) checklist is another tool that was developed for quality appraisal of studies on measurement properties, however, limitations including length of the tool (12 boxes with 119 items), complexity of administration, redundancy of some items and inconsistencies in terminology limit routine use of the tool (Mokkink et al., 2010b, 2010c; Winser et al., 2015). The psychometric properties of other quality assessment tools for validity studies such as the checklist for validation of physical activity instruments (Rennie and Wareham, 1998) and the checklist for evaluating the methodological quality of validation studies on self-report instruments for physical activity and sedentary behavior (Hagströmer, Ainsworth, Kwak, and Bowles, 2012), have not been evaluated (Lucas et al., 2013; Whiting et al., 2005), limiting their applicability as comprehensive tools for quality appraisal of validity studies.

Currently, there are a limited number of tools that can be specifically used to evaluate the quality of validity studies. Therefore, the purpose of this study was to develop a quality appraisal tool specific to validity studies (QAVALS) and to examine its reliability and validity. Developing a tool for quality appraisal of validity studies would serve as a means to evaluate common types of validity of outcome measures assessed in clinical practice and aid in improving the synthesis of information in systematic reviews of the validity of outcome measures. In this manuscript, we describe the development of the Quality Appraisal tool for Validity Studies (QAVALS). The research question this study sought to answer was whether the QAVALS demonstrated evidence of validity and reliability.

Methods

The Institutional Review Board at the University of Michigan–Flint approved this study. The development of the QAVALS followed methods utilized for the design of existing quality appraisal tools including a four-stage process starting with preliminary conceptual decisions, item generation, assessment of content validity, and assessment of reliability (Lucas, Macaskill, Irwig, and Bogduk, 2010; Whiting et al., 2003).

Preliminary conceptual decisions

The preliminary conceptual decisions were based on the design used for the development of the QUADAS and QAREL tools (Lucas, Macaskill, Irwig, and Bogduk, 2010; Whiting et al., 2003). For this study, quality was defined as the extent to which the design, methods, and reporting of the study were in line with the objectives of the study and the extent to which the results of the study were applicable to the target population (Jarde, Losilla, Vives, and Rodrigo, 2013; Lucas, Macaskill, Irwig, and Bogduk, 2010; Whiting et al., 2003). A team of five PhD-trained researchers with backgrounds and expertise in rehabilitation outcomes, health-care epidemiology and statistics was formed for the process of item generation and to provide feedback throughout the process. The researchers had extensive training in research design and methodology, epidemiology, health-care research, statistics and systematic reviews of measurement properties. Based on consensus among the research team, it was agreed that the quality appraisal tool should be able to: 1) Be used to assess the quality of studies included in systematic reviews of validity studies; 2) Be a generic tool that could be used to assess quality of any validity study; 3) Be simple to use and easy to understand; 4) Allow for a reliable assessment of quality by different raters; and 5) Allow for an individual assessment of each item rather than a summary score.

Previous quality appraisal tools have used numeric summary scores for ranking quality of included studies (Lucas, Macaskill, Irwig, and Bogduk, 2010; Lucas et al., 2013). However, the use of summary scores has been questioned in the literature due to problems associated with weighting of individual items (Whiting et al., 2005). Summed quality scores have been shown to differ when items from the same quality appraisal tools were weighted using different methods. Since each item on a quality appraisal tool can individually impact the overall quality of the study, use of summary scores for quality assessment in systematic reviews has been discouraged (Jüni, Witschi, Bloch, and Egger, 1999; Lucas, Macaskill, Irwig, and Bogduk, 2010; Whiting et al., 2005). Based on these observations, it was decided that each item on the QAVALS would be considered separately instead of using an overall quality score.

Item generation

The process of item generation began with an exhaustive review of existing quality appraisal tools for both randomized and non-randomized designs to identify common items that might be relevant for inclusion on the QAVALS. Previously developed rating systems (de Vet et al., 1997; Downs and Black, 1998; Hagströmer,

Ainsworth, Kwak, and Bowles, 2012; McNeely, Olivo, and Magee, 2006; Rennie and Wareham, 1998); quality appraisal tools including the QUADAS 1 and 2 (Whiting et al., 2003, 2011), QAREL (Lucas, Macaskill, Irwig, and Bogduk, 2010), Risk of Bias Assessment tool for Non-randomized designs (RoBANS and RoBINS) (Kim et al., 2013; Sterne et al., 2016), Newcastle–Ottawa scale (Lo, Mertz, and Loeb, 2014; Stang, 2010) and the NIH Quality assessment tool, and quality reporting tools including the STROBE and STARD (Bossuyt et al., 2003; Vandembroucke et al., 2007) were reviewed. The principal investigator mediated feedback between the researchers via online (email/telephonic) or in-person meetings and compiled inputs from each team member regarding potential items to be included. This was then followed by a consensus meeting between research team members to finalize the items that would be included on the initial list. Using the conceptual principles, an initial list of 34 possible items was drafted by the research team for inclusion on the QAVALS (Supplementary File 1).

For each item of the 34 items on the checklist, a detailed list of instructions was developed to aid the rater in the interpretation and scoring of the item and to standardize the rating process (Lucas, Macaskill, Irwig, and Bogduk, 2010). Each item on the tool was designed to be rated on one of the three possible options based on previous rating systems (de Vet et al., 1997; Fuller-Thomson et al., 2009; McNeely, Olivo, and Magee, 2006; National Institutes of Health): ‘Yes’ = meets the criterion, ‘No’ = does not meet the criterion, and ‘Other’ = cannot be determined (CD)/not applicable (NA)/not reported (NR). An item could be rated as CD if the answer to the question could not be determined from the study, as NA if the question was not applicable to the study, and as NR if the information was not reported.

Content validity

A panel of content experts in research design and methodology, statistics, and systematic reviews of measurement properties was invited to establish the content validity of the checklist. The number of panelists chosen was based on previous recommendations (Gilbert and Prion, 2016; Lynn, 1986). A panel of 5–10 experts has been documented to be preferred, and more than 10 experts have been rendered unnecessary for the purpose of content validation (Gilbert and Prion, 2016; Lynn, 1986).

Content validation process round 1

The invited content evaluation panel of 10 experts was provided with the initial version of the QAVALS with 34 items for scoring. They were also provided with an

explanation of the different items on the tool and detailed instructions on scoring of the items. The content validity ratio (CVR) was used for validation. The CVR is widely recognized and one of the most widely used methods for establishing content validity (Gilbert and Prion, 2016; Lynn, 1986; Scally and Ayre, 2014; Wilson, Pan, and Schumsky, 2012). The CVR provides a statistical measure that helps in the decision to reject or retain individual items and has been used extensively in the previous research (Gilbert and Prion, 2016; Lawshe, 1975; Lynn, 1986). Panelists were instructed to independently rate each item on the checklist as “essential”, “useful but not essential”, or “not necessary” according to the criteria originally developed by Lawshe (Gilbert and Prion, 2016; Lawshe, 1975). A period of three weeks was provided to the panelists to complete their ratings, and reminder emails were sent at the end of each week. Individual responses from all the panelists were collected and the number of responses marked as “essential” was identified for each item. For each item, the panelists were also asked to provide reasons for their responses. The CVR for each item was then calculated based on Lawshe’s formula as $CVR = (ne - N/2)/N/2$, where *ne* is the number of panelists identifying an item as “essential” and *N* is the total number of panelists (Lawshe, 1975).

The CVR values range between –1 and +1. CVR values above zero indicate that over half of the panelists agree on an item as “essential”. Lawshe and Schipper’s table of critical values was then used to assess the critical values of CVR ($CVR_{critical}$) to eliminate any chance agreements between experts (Lawshe, 1975). $CVR_{critical}$ is the lowest level of CVR such that, for a given level of significance, the level of agreement exceeds that of chance for a given item (Scally and Ayre, 2014). Items were retained in their original form if the CVRs of the items were above the critical value listed in the table (Lawshe, 1975). All items below the critical CVR value were reviewed and modified (Lawshe, 1975; Scally and Ayre, 2014).

Content validation process round 2

Based on the item CVR and panel feedback, items were either modified or deleted. After modifications to the initial tool were completed, the revised version was sent to the panelists a second time for independent review. Finally, the content validity index (CVI) of the final tool was calculated. CVI is calculated as the mean of the overall CVRs for all items included in the final instrument (Gilbert and Prion, 2016). The CVI provides a numeric value to the content validity of the total scale. A CVI value greater than 0.80 was considered as good content validity (Gilbert and Prion, 2016).

Reliability

Following assessment of content validity, two physical therapists (not involved in the initial development of the QAVALS) with experience in research methods and critical appraisal of studies were invited to participate as raters for reliability testing. Raters were first provided with a trial assessment using a study that was not included in the reliability testing. Following rating of this trial study, raters met with the primary investigator to discuss the criteria for interpretation of each item. After this meeting, each of the two raters was asked to independently rate 10 validity studies. To avoid any form of bias, the raters did not discuss these 10 studies during the trial meeting. The raters were blinded from each other's ratings and were not permitted to discuss their ratings during this process. Both raters were provided with a detailed list of scoring instructions for each item that was developed with the tool.

Since QAVALS was designed as an appraisal tool for systematic reviews that typically evaluate a group of articles of similar content, it was decided to select all 10 studies from validity research on one specific research area for testing reliability. A comprehensive search of PubMed and CINAHL was conducted to locate potential papers on the validity of physical activity monitors. A total of 5392 records were screened to include 25 articles that met the inclusion criteria. Only articles that reported validity of accelerometers were included for this purpose. Of the 25 articles identified, 10 were randomly selected for reliability testing of the QAVALS (Downs and Black, 1998). Instructions on the interpretation of QAVALS items were provided prior to its use.

For test–retest reliability, the raters were asked to rate the same set of 10 studies a second time after a period of 2 weeks (Marx et al., 2003). The raters were advised to destroy their initial ratings after the forms were returned to the primary investigator following the first review to avoid bias.

Analysis

The content validity for the QAVALS was calculated using the CVR and CVI. These scores were used to make decisions regarding item deletion or modification. The cut-off CVR for item deletion was set at 0.50 or if disagreement existed between three or more reviewers. Reliability analysis was performed for the purposes of describing the initial measurement properties of QAVALS, and reliability scores were not utilized to make item deletion or modification decisions. Inter-rater and test–retest reliability of the checklist were assessed using weighted kappa coefficients. Inter-rater reliability was examined first by assessing the overall

agreement between raters on all items (total number of yes, no or other responses that were common between both raters) as well as agreements between raters on individual items. For reliability, kappa coefficient values of 0.8 or more were interpreted as excellent agreement, 0.6–0.8 as good or substantial level of agreement, 0.4–0.6 as moderate, 0.2–0.4 as fair agreement, and values below 0.2 as poor agreement (Portney and Watkins, 2009). Reliability of 0.4 or above was considered as acceptable (Lucas et al., 2013). The level of significance was set at 0.05. All analyses were performed in SPSS version 24.0 (SPSS Inc., Armonk, NY).

Results

Validity outcomes

Eight out of the 10 experts who initially agreed to take part in the process returned completed checklists. One reviewer dropped out due to lack of time and one did not return the first review by the established deadline. Based on the number of experts ($n = 8$), a CVR_{critical} of 0.75 identified from the critical value table, was used to retain items on the tool (Lawshe, 1975). All items that did not reach the critical value of 0.75 were reviewed. Items that were below a CVR of 0.50 or items on which three or more experts disagreed on were eliminated from the checklist. Items that had a CVR of less than 0.75 but more than 0.5 were modified based on the feedback from the experts (Downs and Black, 1998). Following the results of the initial review, five items on the tool were deleted, five items were reframed or modified using the panelist feedback, and four items were combined into two items (Supplementary File 1). The preliminary draft of the QAVALS was modified to have 27 items following the first review. The items that were removed were: 1) Purpose of the study: Panelists felt that since the QAVALS were intended for quality appraisal of validity studies, asking if the purpose of the study was assessing validity was a redundant item; 2) Original source citations: Panelists provided suggestions to combine this item with other items or exclude it. This item as a stand-alone item did not contribute to the quality of validity studies; 3) Unplanned analysis: Panelists suggested that as this item was not addressing the main effects of the study, it was useful to know, but not essential to determine the quality; 4) Generalizability: Panelists thought that individual factors already discussed in the tool were sufficient to determine the generalizability. A separate item was useful but not necessary; and 5) Clinical application: Panelists thought that although this was important, clinical applicability may not always hold true for a validity study as there are studies that validate lab-based instruments.

The revised tool with 27 items was then sent to the panelists a second time. Following the second round, one item on tests of normality was deleted as it had a CVR of 0.25. Based on the feedback received, four additional items were combined to form two items, resulting in the final tool with 24 items. All other items with values above $CVR_{critical}$ were retained. The details on CVR ratings and items modified are listed in Supplementary File 2. The CVI of the items retained on the tool was calculated and was found to be 0.90, which indicated good content validity.

Reliability outcomes

Test-retest reliability for both raters was found to be excellent ($k = 0.84$, 95% CI = 0.76–0.90 for one rater, and $k = 0.80$, 95% CI = 0.76–0.90 for the other rater). The inter-rater reliability of the overall tool was found to be good ($k = 0.70$, 95% CI = 0.61–0.79). When inter-rater reliability of individual items was calculated, it was found to be low for some items and high for some items. Moderate to excellent agreement was observed for seven items (0.41–0.87), fair agreement for two items (0.21–0.34), and poor to no agreement for two items (–0.11 – 0.09). Weighted kappa coefficients for 13 out of the 24 items could not be assessed due to both raters having the same responses to all items resulting from a lack of variability between studies on the items. Table 1 reports the inter-rater reliability of individual items.

The QAVALS

QAVALS consists of 24 questions addressing various aspects of methodological rigor/quality. Each item on the tool can be rated as “yes”, “no”, or “other” and on average, it takes approximately 15 min to rate all the items. The tool is presented in Table 2. A detailed description of each item along with criteria for rating is described in Supplementary File 3.

Discussion

QAVALS was developed using an evidence-based systematic approach to assess the quality of validity studies. The final tool has 24 items where each item can be rated as ‘yes’, ‘no’, or ‘other’. For this study, it was intentionally decided not to use summary scores. Although an overall quantitative score on quality makes the decision-making process easier, problems identified with item weighting have been reported to influence scores (Jüni, Witschi, Bloch, and Egger, 1999; Whiting et al., 2005). Since each item on a quality appraisal tool has its own importance in determining the quality of the study, it is very difficult to find an

objective method to weigh individual items on a scale as the criteria for weighting of items are usually subjective and arbitrary (Whiting et al., 2005). Using summary scores in a systematic review can lead to different conclusions, thereby affecting the overall quality of the review (Whiting et al., 2005).

For this tool, an overall subjective quality grading (good, fair, poor or low, moderate, high risk of bias) was not used to determine the study quality. Variable responses were received when content experts were asked to provide their opinions regarding inclusion of an overall subjective grading. Although most experts thought that a subjective quality grading would be useful to the reader and more informative, there was mutual consensus that inclusion of quality grading would include the possibility of ambiguity in the rating. Another problem with inclusion of overall quality grading was the development of criteria to establish quality. A straightforward method would be to count the number of ‘yes’ responses and establish a cut off for ‘yes’ responses beyond which the study would qualify as a good quality study. However, use of this method defies the very concept of not using summary scores and would automatically weight each item evenly. The other problem with the use of this method was that it would only consider the ‘yes’ responses to establish whether a study was of good quality. Since QAVALS has several items that could be rated as ‘not applicable’, a greater number of ‘NA’ items may inadvertently bring down the quality of the study irrespective of design. Another way to approach this problem would be to have the rater use his discretion based on his rating of individual items to grade quality. However, since these ratings would then be highly subjective and based on the rater’s perspective of quality, the tool would become highly specific to the use of experienced raters or would require a high level of training before use of this tool.

Different cut-offs for CVI values have been reported in the literature as criteria for establishing good content validity ranging from 0.70 (Tilden, Nelson, and May, 1990) to as high as 0.80 (Davis, 1992). The CVI for QAVALS was found to be 0.90 in this study which is considered as evidence of strong content validity.

Overall, the QAVALS tool was found to have excellent test-retest and good inter-rater reliability. Although four items showed fair to poor agreement between raters, most items demonstrated moderate to good agreement. (Table 1). Based on these results, the QAVALS is considered a sufficiently reliable tool to assess the quality of studies of validity.

The two items that demonstrated poor agreement between raters were items describing the outcomes to be validated and the procedures for testing validity. Possible explanation for the low reliability of these items could be explained by the way these were structured, creating

Table 1. Inter-rater reliability of individual items on the QAVALS.

QAVALS item	Weighted Kappa	p-Value (95% CI)
1 Was the study design reported?	-	
2 Did the study provide an accurate description of the type of validity tested?	0.54	0.53 (0.44–1.04)
3 Was the study setting and time frame of participant recruitment clearly outlined and described?	0.44	0.48 (–0.2–1.08)
4 Were the criteria for participant selection clearly described?	0.87	0.001 (0.62–1.12)
5 Were the participants in the study representative of the sample population from which they were recruited?	-	
6 Did the study clearly describe the outcome measures to be validated?	–0.11	0.72 (–0.26–0.42)
7 Did the study provide a clear description of the procedures for testing validity?	0.09	0.87 (–0.47–0.65)
8 Was the testing procedure standardized for all participants?	0.34	0.08 (–0.03–0.71)
9 Was a priori sample size calculation performed to ensure that the study had sufficient power?	0.60	0.03 (0.14–1.05)
10 Did the study describe and justify any attrition that may have occurred?	-	
11 Were the statistical analyses used to test validity appropriate for the study?	-	
12 When multiple comparisons were performed, were appropriate statistical adjustments used to control for the likelihood of a type 1 error?	0.41	0.10 (–0.18–1.00)
13 Did the study identify potential confounding variables and if so, were measures taken to adjust for these confounders?	-	
14 Were the primary findings of the study clearly described?	-	
15 Were validity coefficients reported for primary outcomes?	-	
16 For primary outcomes, did the study report the standard deviation or confidence intervals for normally distributed data? Or, if non-normally distributed data, did the study report the inter-quartile range for the main outcomes?	-	
17 Was the process of selecting expert panel and their qualifications described?	-	
18 Did the study provide a rationale for the selection of the reference standard?	0.48	0.04 (–0.04–1.01)
19 When the index test was assessed by more than one rater, were the raters blinded to the findings of the other raters?	-	
20 When the index test was assessed by more than one rater, was the inter-rater reliability between raters established and reported?	-	
21 Was the time interval used between administration of reference standard and the test measure appropriate?	0.61	0.03 (–0.04–1.27)
22 Were subjects in different groups homogenous at baseline or if they weren't homogenous at baseline, were differences between groups accounted for during the analysis?	-	
23 Did the measures used for convergent validity represent a similar construct as the outcome measure of interest?	0.21	0.49 (–0.43–0.85)
24 Did the measures used for discriminant validity represent a construct different from the outcome measure of interest?	-	

CI = confidence interval; '-' = the k statistic could not be calculated because both raters had same responses.

a degree of ambiguity in these items. For example, in item 6: “Did the study clearly describe the outcome measures to be validated?” and item 7: “Did the study provide a clear description of the procedures for testing validity?”, the use of the word “clearly” made these questions open to interpretation. What one rater considers as a ‘clear description’ might not be the same for another rater. Additionally, fair reliability noted on two of the items could be attributed to the lack of clarity in the detailed description of these items. For item 8 “Was the testing procedure standardized for all participants”, the detailed explanation (Supplementary File 3) was that all participants would have received testing in the same manner. However, the description of the item did not include instructions on the order of examination which may have led to differences in rating. Raters could consider that the testing was standardized if the participants were tested using the same instruments irrespective of the order of examination. Similarly, for item 23: “Did the measures used for convergent validity represent a similar construct as the outcome measure of interest?”, the description of the item did not specify that the measures used should be valid for the evaluation of the construct in question. Not using

the term ‘validity’ in the description may not have provided a clear direction to the raters.

Future research to examine the reliability of QAVALS using studies on different outcome measures should be performed. Weighted kappa coefficients could not be calculated for 13 items on this tool where both raters gave the same responses across all studies. This was because of a lack of variability between the studies on these response categories (Chmura Kraemer, Periyakoil, and Noda, 2002). Since weighted kappa coefficients compare the variability between pairs of items to the total variability across studies, low variance between studies may result in large error variance in relation to the study variance and hence, this measurement cannot be performed (Chmura Kraemer, Periyakoil, and Noda, 2002; Mokkink et al., 2010a). However, it is encouraging that both raters gave the same responses across all studies for those items. Future studies utilizing multiple examiners and more studies to explore the inter-rater reliability may potentially help in generating reliability values for the remaining items.

Table 2. Quality assessment of validity studies (QAVALS).

Item	Item criteria			Other (CD, NR,
		Yes	No	NA*
1	Was the study design reported?			
2	Did the study provide an accurate description of the type of validity tested?			
3	Was the study setting and time frame of participant recruitment clearly described?			
4	Were the criteria for participant selection clearly described?			
5	Were the participants in the study representative of the sample population from which they were recruited?			
6	Did the study clearly describe the outcome measures to be validated?			
7	Did the study provide a clear description of the procedures for testing validity?			
8	Was the testing procedure standardized for all participants?			
9	Was a priori sample size calculation performed to ensure that the study had sufficient power?			
10	Did the study describe and justify any attrition that may have occurred?			
11	Were statistical analyses used to test validity appropriate for the study?			
12	When multiple comparisons were performed, were appropriate statistical adjustments used to control for the likelihood of a type 1 error?			
13	Did the study identify potential confounding variables and if so, were measures taken to adjust for these confounders?			
14	Were primary findings of the study clearly described?			
15	Were validity coefficients reported for primary outcomes?			
16	For primary outcomes, did the study report standard deviations or confidence intervals for normally distributed data? If non-normally distributed data, did the study report inter-quartile ranges for the main outcomes?			
<i>Face and content validity:</i>				
17	Was the process of selecting expert panel and their qualifications described?			
<i>Criterion validity:</i>				
18	Did the study provide a rationale for the selection of the reference standard?			
19	When the index test was assessed by more than one rater, were the raters blinded to the findings of the other raters?			
20	When the index test was assessed by more than one rater, was the inter-rater reliability between raters established and reported?			
21	Was the time interval used between administration of reference standard and the test measure appropriate?			
<i>Construct validity (known groups):</i>				
22	Were subjects in different groups homogenous at baseline? If they weren't homogenous at baseline, were differences between groups accounted for during the analysis?			
<i>Construct validity (convergent):</i>				
23	Did the measures used for convergent validity represent a similar construct as the outcome measure of interest?			
<i>Construct validity (discriminant):</i>				
24	Did the measures used for discriminant validity represent a construct different from the outcome measure of interest?			

*CD = cannot be determined; NA = not applicable; NR = not reported.

The studies included for reliability testing in this study were identified from a previous systematic review performed. These studies were selected to limit the studies to one area of interest, making it consistent with the usual systematic review quality appraisal process. Lucas et al. (2013) indicated that using studies of similar topic is a preferred method for reliability testing of quality appraisal studies. On the other hand, it has been noted that limiting studies to only one area of diagnostic technology may result in low inter-rater reliability (Hollingworth et al., 2006). However, this was not observed in our study.

Limitations

QAVALS was limited in its ability to distinguish between reporting quality and methodological quality. The quality of a study strongly depends not only on the design of methods but also on the reporting of methods and results. Additionally, although the final version was formed after removing several items from the original pool, the tool still had 24 items and was considerably lengthy. Future studies to develop a shorter version of this checklist may be helpful. Since the reliability testing was performed on a limited number of studies using only two raters, future work is required to test the

reliability further using a larger number of studies and multiple raters with expertise in varied areas of research to potentially concise the tool.

This tool was designed to serve as a starting point for quality appraisal of common validity types. Future studies should focus on development at tools for advance validation methods including factor analysis and structural equation modeling.

Finally, it was found that the raters had difficulty understanding the distinction between 'unclear' and 'cannot be determined' responses and more clarity on these responses would aid in a better rating. Since these responses were part of a single rating category of 'other', the difference in responses within this category did not affect the overall reliability of the tool. However, a shorter version of this tool in future may be developed with the use of a single response category. The exploration of the concurrent validity of QAVALS against other quality appraisal tools is also recommended.

Conclusion

This study presents a new valid and reliable tool for quality appraisal of validity studies and can be used in the risk of bias assessment of included studies in

systematic reviews of validation studies. This tool includes descriptors for each item and self-explanatory scoring instructions that require no additional training. Future research to examine its concurrent validity against other quality appraisal tools as well as to explore its reliability further using multiple raters is recommended.

Declaration of Interest

The authors declare no conflict of interest.

ORCID

Shweta Gore  <http://orcid.org/0000-0002-5802-182X>

References

- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG **2003** The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry* 49: 7–18.
- Chmura Kraemer H, Periyakoil VS, Noda A **2002** Kappa coefficients in medical research. *Statistics in Medicine* 21: 2109–2129.
- Crowe M, Sheppard L **2011** A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed. *Journal of Clinical Epidemiology* 64: 79–89.
- Davis LL **1992** Instrument review: Getting the most from a panel of experts. *Applied Nursing Research* 5: 194–197.
- de Vet HC, de Bie RA, van der Heijden GJ, Verhagen AP, Sijpkens P, Knipschild PG **1997** Systematic reviews on the basis of methodological criteria. *Physiotherapy* 83: 284–289.
- Downs SH, Black N **1998** The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health* 52: 377–384.
- Drucker AM, Fleming P, Chan AW **2016** Research techniques made simple: Assessing risk of bias in systematic reviews. *Journal of Investigative Dermatology* 136: e109–e114.
- Fuller-Thomson E, Yu B, Nuru-Jeter A, Guralnik JM, Minkler M **2009** Basic ADL disability and functional limitation rates among older americans from 2000-2005: The end of the decline? *Journals of Gerontology Series A* 64A: 1333–1336.
- Gilbert GE, Prion S **2016** Making sense of methods and measurement: Lawshe's content validity index. *Clinical Simulation in Nursing* 12: 530–531.
- Gopalakrishnan S, Ganeshkumar P **2013** Systematic reviews and meta-analysis: Understanding the best evidence in primary healthcare. *Journal of Family Medicine and Primary Care* 2: 9–14.
- Hagströmer M, Ainsworth BE, Kwak L, Bowles HR **2012** A checklist for evaluating the methodological quality of validation studies on self-report instruments for physical activity and sedentary behavior. *Journal of Physical Activity and Health* 9: S29–S36.
- Hollingworth W, Medina LS, Lenkinski RE, Shibata DK, Bernal B, Zurakowski D, Comstock B, Jarvik JG **2006** Interrater reliability in assessing quality of diagnostic accuracy studies using the QUADAS tool. A Preliminary Assessment. *Academic Radiology* 13: 803–810.
- Institute of Medicine (US) Committee on Standards for Systematic Reviews of Comparative Effectiveness Research **2011** Finding what works in health care: Standards for systematic reviews. Washington, D.C: National Academies Press.
- Jarde A, Losilla J, Vives J, Rodrigo FM **2013** Q-Coh: A tool to screen the methodological quality of cohort studies in systematic reviews and meta-analyses. *International Journal of Clinical and Health Psychology* 13: 138–146.
- Jüni P, Witschi A, Bloch R, Egger M **1999** The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 282: 1054–1060.
- Kim SY, Park JE, Lee YJ, Seo H-J, Sheen SS, Hahn S, Jang BH, Son HJ **2013** Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of Clinical Epidemiology* 66: 408–414.
- Lawshe CH **1975** A quantitative approach to content validity. *Personnel Psychology* 28: 563–575.
- Lo CK, Mertz D, Loeb M **2014** Newcastle-Ottawa Scale: Comparing reviewers' to authors' assessments. *BioMed Central Medical Research Methodology* 14: 45.
- Lucas N, Macaskill P, Irwig L, Bogduk N **2010** The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology* 63: 854–861.
- Lucas N, Macaskill P, Irwig L, Moran R, Rickards L, Turner R, Bogduk N **2013** The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BioMed Central Medical Research Methodology* 13: 111.
- Lynn MR **1986** Determination and quantification of content validity. *Nursing Research* 35: 382–386.
- Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF **2003** A comparison of two time intervals for test-retest reliability of health status instruments. *Journal of Clinical Epidemiology* 56: 730–735.
- McNeely ML, Olivo SA, Magee DJ **2006** A systematic review of the effectiveness of physical therapy interventions for temporomandibular disorders. *Physical Therapy* 86: 710–725.
- Mokkink L, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, Knol DL, Bouter LM, de Vet HC **2010a** Interrater agreement and reliability of the COSMIN (Consensus-based Standards for the selection of health status Measurement Instruments) checklist. *BioMed Central Medical Research Methodology* 10: 82.
- Mokkink L, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HC **2010b** The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. *BioMed Central Medical Research Methodology* 10: 22.
- Mokkink L, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC **2010c** The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international delphi study. *Quality of Life Research* 19: 539–549.
- Portney L, Watkins M **2009** Foundations of clinical research: Applications to practice (3rd). Philadelphia, F.A: Davis Company.

- Rennie KL, Wareham NJ 1998 The validation of physical activity instruments for measuring energy expenditure: Problems and pitfalls. *Public Health Nutrition* 1: 265–271.
- Scally AJ, Ayre C 2014 Critical values for lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development* 47: 79–86.
- Stang A 2010 Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. *European Journal of Epidemiology* 25: 603–605.
- Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I et al. 2016 ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *British Medical Journal* 355: i4919.
- Tilden VP, Nelson CA, May BA 1990 Use of qualitative methods to enhance content validity. *Nursing Research* 39: 172–175.
- Vandenbroucke JP, Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, Poole C, Schlesselman JJ, Egger M 2007 Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. *Epidemiology* 18: 805–835.
- Whiting P, Rutjes AW, Dinnes J, Reitsma JB, Bossuyt PM, Kleijnen J 2005 A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools. *Journal of Clinical Epidemiology* 58: 1–12.
- Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J 2003 The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BioMed Central Medical Research Methodology* 3: 25.
- Whiting P, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, Altman D et al. 2011 QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 155: 529–536.
- Wilson FR, Pan W, Schumsky DA 2012 Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development* 45: 197–210.
- Winser SJ, Smith CM, Hale LA, Claydon LS, Whitney SL, Mehta P 2015 COSMIN for quality rating systematic reviews on psychometric properties. *Physical Therapy Reviews* 20: 132–134.