Factor Analysis Methods and Validity Evidence: A Review of Instrument Development Across the Medical Education Continuum

Angela P. Wetzel, PhD

Abstract

Purpose
Instrument development consistent with best practices is necessary for effective assessment and evaluation of learners and programs across the medical education continuum. The author explored the extent to which current factor analytic methods and other techniques for establishing validity are consistent with best practices.

Method
The author conducted electronic and hand searches of the English-language medical education literature published January 2006 through December 2010. To describe and assess current practices, she systematically abstracted reliability and validity evidence as well as factor analysis methods, data analysis, and reported evidence from instrument development articles reporting the application of exploratory factor analysis and principal component analysis.

Results
Sixty-two articles met eligibility criteria. They described 64 instruments and 95 factor analyses. Most studies provided at least one source of evidence based on test content. Almost all reported internal consistency, providing evidence based on internal structure. Evidence based on response process and relationships with other variables was reported less often, and evidence based on consequences of testing was not identified. Factor analysis findings suggest common method selection errors and critical omissions in reporting.

Conclusions
Given the limited reliability and validity evidence provided for the reviewed instruments, educators should carefully consider the available supporting evidence before adopting and applying published instruments. Researchers should design for, test, and report additional evidence to strengthen the argument for reliability and validity of these measures for research and practice.

Across the continuum of medical education, written examinations, questionnaires, performance-based checklists, and objective structured clinical examinations are used to assess outcomes at levels ranging from the learner to the individual patient and, ultimately, to the health of the community.[1] These instruments must be developed carefully if they are to measure outcomes precisely and accurately. If they are poorly designed, there is an increased risk that they will lead to misinformed or inaccurate conclusions about learner knowledge, skills, and attitudes, or program effectiveness. The potential impact of this risk depends on the proposed use of the instrument's scores. For example, a poorly designed instrument could result in inaccurate formative feedback to a third-year clerk, an inaccurate statement of a resident's competency at a surgical procedure, or a misinformed decision to reallocate resources and terminate a program. Therefore, it is critical that the quality of measurement be consistent with best practices for reliability and validity evidence, especially for high-stakes summative assessments and for credible program evaluation.

Factor analysis is one method that is useful for establishing evidence for validity.[2] Yet, psychology and general education literature reviews[2–8] of factor analysis for instrument development suggest methodological errors and omissions in reporting, thus limiting the potential for evaluation and replication. In the medical education literature, more broadly focused reviews[9–16] consider multiple sources of reliability and validity evidence in instrument development; however, insufficient reporting similarly limits the ability of medical educators and researchers to evaluate instruments for use. Existing studies of measures focus on select topics such as professionalism,[11,16] script concordance,[12] and continuing medical education,[13] but, to the best of my knowledge, there has not been a comprehensive review of instrument development across the continuum of medical education.

To address that gap, I reviewed medical education (undergraduate, graduate, and continuing) instrument development articles that report exploratory factor analysis (EFA) or principal component analysis (PCA) to describe and assess their reliability and validity evidence, including factor analysis. Findings from this study inform two research questions: Within the medical education instrument development literature, (1) to what extent are techniques for establishing validity consistent with the *Standards for Educational and Psychological Testing*,[17] and (2) to what extent are EFA and PCA methods, data analysis, and reported evidence consistent with factor analytic best practices?

Dr. Wetzel is director of assessment, Department of Foundations of Education, Virginia Commonwealth University School of Education, Richmond, Virginia. At the time of writing, she was a graduate assistant, Office of Assessment and Evaluation Studies, Virginia Commonwealth University School of Medicine, Richmond, Virginia.

Correspondence should be addressed to Dr. Wetzel, Office of Assessment, Department of Foundations of Education, VCU School of Education, 1015 West Main St., Room 2128, Box 842020, Richmond, VA 23284-2020; telephone: (804) 828-8673; e-mail: apwetzel@vcu.edu.

Acad Med. 2012;87:1060–1069.
First published online June 20, 2012
doi: 10.1097/ACM.0b013e31825d305d

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A98.

Method

Literature search and eligibility criteria

I conducted an electronic search of the medical education literature in the MEDLINE, ERIC, PsycINFO, and CINAHL databases, using variations of the following search terms as they appear in the thesaurus for each database: *validity, reliability, test construction, psychometrics, factor analysis, measures (individuals), measurement, medical school, medical education, medical student.* All medical education research articles that met the following criteria were included in the review: (1) human study, including but not limited to medical students, residents, or physicians, (2) development of a new or revised instrument measuring knowledge, skills, or attitudes or medical education program effectiveness, (3) application of EFA or PCA, (4) published in English, and (5) published from January 2006 through December 2010. I reviewed titles, abstracts, and full text as needed to determine fit for inclusion according to these eligibility criteria. Lastly, I hand searched the reference lists of all included articles.

Data abstraction form development

I developed a data abstraction form and coding manual, informed by best practices derived from the literature,[17–25] and pilot tested them with five sample articles. An additional trained coder participated in a second pilot test using an additional five sample articles. These pilot tests identified revisions for the form and manual to improve coding consistency and data quality.

The final coding manual and data abstraction form included four sections: (1) descriptive information about the article (e.g., journal, construct measured), (2) educational outcome level (e.g., satisfaction, competence),[1] (3) factor analysis methods (e.g., extraction method, criteria for factor retention), and (4) other techniques for establishing validity evidence (e.g., reliability measures, expert review, predictive or concurrent criterion validity). Sections two through four consisted of dichotomous check boxes for indicating which outcome levels, factor analysis methods, and other validity techniques were present in each article. I used the form and manual to systematically abstract data from all articles selected for inclusion in the review. The second reviewer coded six randomly selected articles from the final set in a peer-review process. As coding decision points were dichotomous check boxes, agreement occurred when we both consistently indicated a characteristic as present or not present in the study. We discussed disagreements until we reached consensus. The calculated agreement for these six articles using proportion of total agreements was 93.4% (range: 80.9%–100%).

Data abstraction and synthesis

The data abstraction process began with documenting descriptive information for each article, including coding the outcome assessed or evaluated by the study instrument using Moore and colleagues'[1] outcomes framework for participant satisfaction, declarative and procedural knowledge, competence, performance, patient health, and community health. Next, I abstracted specifics related to factor analysis methods using a framework of best practices derived from the literature[18–24]: sample size criteria, model of analysis, extraction and rotation method, criteria for factor retention, and factor loadings.

Finally, I coded each article for the other techniques the researchers applied for establishing validity evidence. Historically, instrument validation included efforts to investigate three distinct types of validity—content, criterion, and construct validity—to establish a measure as reliable and valid. Conceptual changes in the measurement field, however, emphasize that reliability and validity are not inherent to an instrument but, rather, represent an interaction between the measure, the setting, and the sample.[26–28] A contemporary perspective emerged, with recommendations for best practices in *Standards for Educational and Psychological Testing*[17] asserting validity as a contextually specific and unitary concept supported by accumulated evidence from five sources: test content, response process, internal structure, relationships with other variables, and consequences of testing. Yet, traditional terminology associated with validity types remains in active use in medical education.[10,16,25,29] As such, I abstracted types of reliability and validity as reported in the articles. To illustrate current practices in relation to contemporary best practices,[17] I mapped the traditional approaches onto the contemporary framework for interpretation. A comparison of the factor analysis methods and other validity evidence to contemporary best practices enabled evaluation of current practices.

Results

Of the 907 articles identified through electronic and hand searches, 62 met the eligibility criteria after accounting for duplicates (Figure 1).[30–91] Almost all of the included articles (n = 60; 96.8%) discussed the development of one instrument, whereas two (3.2%) discussed the development of two instruments, resulting in a total of 64 instruments reviewed. Fourteen articles (22.6%) included more than one factor analysis; I coded each of these analyses individually for a total of 95 factor analyses reviewed. Results are reported in frequency tables to provide a descriptive summary of current instrument development practices in medical education.

Table 1 describes publication characteristics as well as constructs measured and educational outcome levels assessed. (Article-level details about respondent type and instruments used in the studies are provided in Supplemental Digital Appendix 1, http://links.lww.com/ACADMED/A98.)

Techniques for establishing validity evidence

All techniques reported for establishing reliability and validity evidence are detailed in Table 2. Results are described using traditional validity terminology and are presented according to contemporary sources of validity evidence.[17]

Evidence based on test content. For 44 (68.8%) of the 64 instruments, researchers reported evidence consistent with a traditional definition of content validity, including item development based on a review of the literature (n = 25), review of items by a sample from the target population (n = 16), and use of previously tested items (n = 9). However, using the contemporary framework, best practices include reporting three key sources of evidence based on test content—traditional content validity plus expert review and pilot testing.[17]

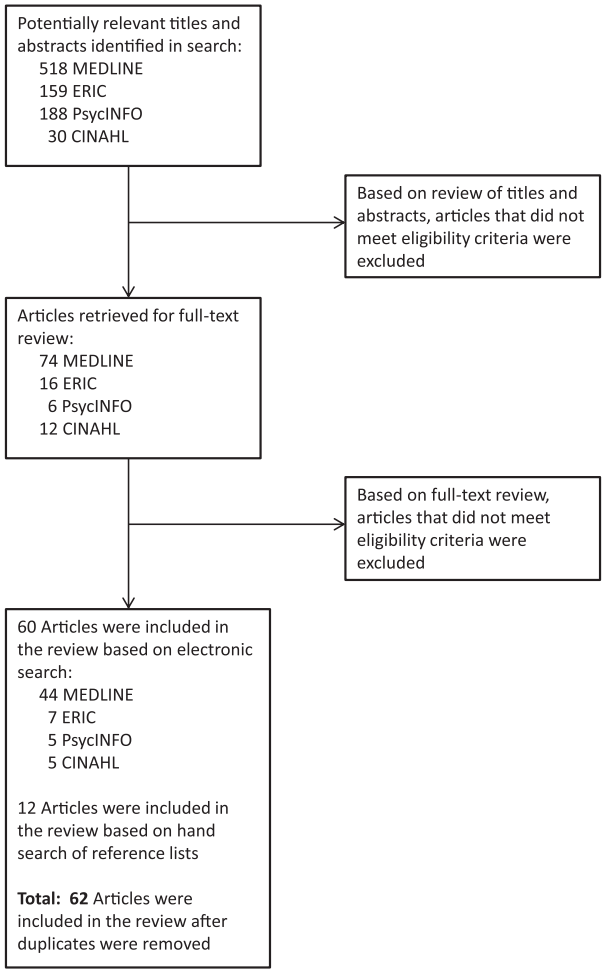

Figure 1 Literature search details. Categories may not be mutually exclusive.

From this perspective, only 9 (14.1%) of the instruments were supported with all three endorsed sources of evidence; 23 (35.9%) were supported with one of these sources and 17 (26.6%) with two. Authors employed expert review of items for 24 (37.5%) of the 64 instruments; 19 of these (29.7% of 64) were accompanied by full description of the qualifications of the experts and the process of review. Pilot testing with the target population occurred for 16 (25.0%) of the instruments. Face validity—a term no longer supported in the contemporary perspective—was reported as supportive evidence for 11 instruments (17.2%).

Evidence based on relationships with other variables. *Concurrent* and *predictive criterion validity* and *convergent*, *discriminant*, and *divergent validity* are traditional terms related to validity evidence based on relationships with other variables. Of these, investigators most frequently reported divergent validity evidence (n = 25/64; 39.1%). Predictive criterion evidence was not reported for any instrument.

Evidence based on response process. Findings related to evidence based on response process are presented in Table 2. Interrater and intrarater reliability were only relevant to six instruments (9.4%) in studies that involved multiple raters per evaluand or multiple evaluands per individual rater. Of these instruments, interrater reliability was reported for three (50.0%); intrarater reliability was reported for none. As none of the instruments had multiple forms, investigators did not report alternate-forms reliability.

Evidence based on internal structure. All studies in this review employed factor analysis; therefore, reporting for all 64 instruments included evidence based on dimensionality to support internal structure. Researchers reported evidence for internal consistency reliability for almost all (n = 59; 92.2%) reviewed instruments. Of the seven single-dimension instruments, internal consistency reliability was calculated for six (85.7%); researchers reported at the total scale level for all six (100%). Of the 57 multidimensional instruments, internal consistency reliability was calculated for 53 (93%). Among these, internal consistency was calculated at the total scale level for 16 (30.2%), at the subscale level for 21 (39.6%), and at both the total scale and subscale levels for 16 (30.2%). When estimating internal consistency, investigators most often applied Cronbach alpha. Some researchers used item–scale and item–total correlations and reliability-if-item-deleted to determine which items to retain.

Evidence based on consequences of testing. Researchers did not report evidence based on consequences of testing for any of the 64 instruments.

Factor analysis methods

All 95 factor analysis methods reported in the studies reviewed are presented in Table 3.

Sample size. Sample sizes used to run the 95 factor analyses ranged from fewer than 100 respondents (n = 13; 13.7%) to more than 500 respondents (n = 13; 13.7%). Sixty-two (65.3%) had 300 or fewer respondents. Of the 87 factor analyses that reported sample size, 83 (95.4%) provided the total number of items in the final instrument, allowing calculation of the participant-to-item ratio. This value ranged from 1.54:1 to 3140.45:1, with a mean of 55.7:1 and a median of 11.55:1; 46 analyses (55.4%) met or exceeded a 10:1 ratio.

Model of analysis and extraction method. Among the 95 factor analyses, PCA was the most frequently applied model and extraction method (n = 60; 63.2%). Investigators described 35 of the analyses as EFA, yet I determined that 19 (54.3%) of these 35 were PCA. Further, researchers incorrectly reported use of confirmatory factor analysis (CFA) for three additional analyses for which the researchers applied EFA. I found that just

Table 1
Characteristics of the 62 Medical Education Instrument Development Articles Employing Factor Analysis and Their 64 Instruments

Characteristic	Number of articles or instruments	%
Year of publication (n = 62)		
2006	10	16.1
2007	10	16.1
2008	9	14.5
2009	22	35.5
2010	11	17.7
Journal (n = 62)		
Medical Teacher	13	21.0
Academic Medicine	10	16.1
Medical Education	5	8.1
Journal of General Internal Medicine	5	8.1
Advances in Health Sciences Education	2	3.2
Education for Health	2	3.2
Patient Education and Counseling	2	3.2
Other*	23	37.1
Construct measured by instrument (n = 64)		
Clinical specific knowledge, skill, or attitude	10	15.6
Career preference	7	10.9
Professionalism	7	10.9
Educational environment	5	7.8
Instructional quality	5	7.8
Communication and feedback skills	5	7.8
Self-directed/lifelong learning	4	6.3
Empathy	4	6.3
Learning style, behavior, or skill	4	6.3
Interprofessional teams, teams, and team leadership	3	6.3
Patient safety	2	3.1
Educational program quality	2	3.1
Miscellaneous	6	9.4
Educational outcome level assessed† by instrument (n = 64)		
Level 2: Participant satisfaction	13	20.3
Level 3A: Declarative knowledge/attitude	36	56.3
Level 3B: Procedural knowledge/attitude	0	0
Level 4: Competence	4	6.3
Level 5: Performance	8	12.5
Level 6: Patient health	0	0
Level 7: Community health	0	0
Unclear	3	4.7

*Includes 23 journals that each published 1 article included in this review.
† Educational outcome levels derived from Moore and colleagues.[1]

16 (16.8%) of the 95 analyses appropriately employed an exploratory factor model (see Table 3 for extraction methods used).

Rotation method. In total, 62 (65.3%) of the 95 factor analyses interpreted an orthogonal rotation for the factor solution, including 7 (7.4%) that first explored both orthogonal and oblique factor rotations. Fewer analyses (n = 20; 21.1%) interpreted an oblique rotation. Reporting for only 25 analyses (26.3%) included justification for the selection of a rotation method based on evidence for the relationships between factors.

Criteria for factor retention. Overall, 42 (44.2%) of the 95 factor analyses applied one criterion in determining the number of factors to retain, 30 (31.6%) used two criteria, and 12 (12.6%) considered three or more criteria. The remaining 11 (11.6%) failed to report which criteria were used. The most frequently applied criteria included the Kaiser criterion (n = 46; 48.4%),[92] Cattell scree test (n = 35; 33.7%),[93] conceptual meaningfulness of each factor (n = 21; 22.1%),[19] and minimum number of items required per factor (n = 18; 18.9%).[20,24]

Factor loadings. Thirty-three (34.7%) of the 95 factor analyses included all factor loadings for all items, making clear to the reader the distribution of items across factors. For example, Carruthers and colleagues[34] reported in their study that the item "All medical errors should be reported" loaded on the factor named "disclosure responsibility." Thirty (31.6%) analyses reported only factor loadings for items that met a certain criterion, but 32 (33.7%) reported no factor loadings.

Discussion

The findings of this review indicate a tendency among medical education researchers to report validity evidence based on test content and internal structure and to exclude investigation of other evidence, including that based on response process, relationships with other variables, and consequences of testing. Findings related to factor analysis current practices suggest common errors in selecting factor analysis methods and in reporting evidence. Further, critical omissions in reporting of information limit the potential for replication and verification by other researchers and the evaluation by educators who may seek to apply the instrument in their practice.

Validity evidence

This review provides evidence that investigators retain the traditional validity framework to support medical education instrument development. For instance, a number of authors suggested that their findings established an instrument's construct validity. However, from a contemporary

Table 2

Reported Evidence for Reliability and Validity of 64 Medical Education Instruments Abstracted Using a Traditional Validity Framework and Mapped to the Contemporary Framework[17] of Validity as a Unitary Concept

Type of evidence	Number of instruments	%
Evidence based on test content*		
Face validity	11	17.2
Content validity	44	68.8
Expert review	24	37.5
Pilot test	16	25.0
Evidence based on relationships with other variables*		
Concurrent criterion validity	6	9.4
Predictive criterion validity	0	0
Convergent evidence	8	12.5
Discriminant evidence	1	1.6
Divergent evidence	25	39.1
Evidence based on response process		
Intrarater reliability†	0	0
Interrater reliability†	3	50.0
Test–retest reliability*	4	6.3
Test–retest stability*	4	6.3
Questioning test takers about process of response to items (i.e., cognitive interviewing)*	5	7.8
Generalizability theory*	4	6.3
Evidence based on internal structure*		
Internal consistency	59	92.2
Alternative-form reliability	0	0

* n = 64.
† Potential n = 6.

perspective,[17] all validity evidence supports construct validity; therefore, the term *construct validity* did not always convey substantial meaning or communicate which techniques the study authors applied for establishing validity. Researchers made infrequent references to language from the contemporary sources of validity evidence (e.g., evidence based on internal structure,[59,61,90] evidence based on test content[59,90]). It is unclear why the transition from the traditional to the contemporary validity framework, which was introduced in 1999, has yet to occur in medical education. It is necessary, however, to discard traditional notions of validity types and replace them with contemporary best practices that emphasize quality instrument development through rigorous reliability and validity testing across time, settings, and samples in order to build evidence, supported by multiple sources, for a measure's intended use.[9,13,14,16,94] Although most instruments included some evidence based on test content, less than 15% of reviewed instruments included all three recommended elements (i.e., traditional content validity, expert review of items, pilot testing). In 20% of the articles reporting that expert review was employed, authors did not fully describe the qualifications of the experts and the process of review. Pilot testing, which occurred for just 25% of the instruments, can present feasibility challenges, particularly in studies where access to participants is limited. To the extent possible, though, pilot testing or at least review of potential items by a subset of the target population is highly preferred to ensure clarity and relevance of the items prior to administration.[19,25]

Empirical analysis to examine the underlying dimensions of a new measure is important, and researchers did conduct variations of factor analysis in the reviewed studies; however, conducting an EFA is not, on its own, sufficient evidence for internal structure. The researcher must establish, for the reader, the link between the empirically derived factor structure and the structure of the construct informed by the literature. For example, Donnon and colleagues[40] made clear the relationships between the seven factors retained for their Rural Integrated Community Clerkship questionnaire and the key themes that emerged from student interviews during the item development process. Researchers did not always include this additional step in the studies reviewed, which made it difficult to translate what the factor analysis and evidence for multiple dimensions added as supportive evidence, if anything.

Following an EFA, instrument development should include calculation of internal consistency,[17] and 92% of the investigators reported this evidence for the total scale, the subscales, or both. Cronbach alpha was most often used, yet it is not necessarily appropriate for all internal consistency calculations. Specifically, summation of total scores is not appropriate for multidimensional instruments; therefore, Cronbach alpha should be limited to subscales[95] as demonstrated in 40% of the multidimensional instruments in this review. The omega reliability statistic resolves the issues of alpha and provides a means of calculating a more precise measure of internal consistency for subscales and total scales for multidimensional instruments.[95] The use of omega was not identified in this review, and the statistical calculation is not available in common social science statistical software.

Although individual measures of reliability rule out threats based on specific sources (e.g., time or multiple raters), reporting of multiple reliability measures best supports the argument for reliability of an instrument.[17] Further, generalizability theory applies a random analysis of variance model to test the influence of multiple factors on the reliability of an instrument. Although generalizability theory was applied in several of the reviewed studies, its statistical assumptions

Table 3

Methods and Reporting of the 95 Factor Analyses Applied in the 62 Medical Education Instrument Development Articles Reviewed

Factor analysis methods and reporting	Number of factor analyses	% of 95
Sample size		
≤100	13	13.7
101–200	25	26.3
201–300	24	25.3
301–400	9	9.5
401–500	3	3.2
≥501	13	13.7
Unclear	8	8.4
Extraction method		
Principal components analysis (PCA)	60	63.2
Exploratory factor analysis (EFA)*	16	16.8
Principal axis factoring (PAF)	4	4.2
Maximum likelihood	8	8.4
Weighted least squares	1	1.1
Unweighted least squares	2	2.1
Combination of PCA and PAF†	1	1.1
Not reported	16	16.8
Unclear	3	3.2
Rotation method		
Orthogonal‡	55	57.9
Varimax	61	64.2
Not reported	1	1.1
Oblique‡	20	21.1
Promax	7	7.4
Direct oblimin	17	17.9
Not reported	2	2.1
Unclear	1	1.1
Combination orthogonal and oblique§	7	7.4
No rotation	1	1.1
Not reported	10	10.5
Unclear	2	2.1
Criteria for factor retention		
Previous theory	4	4.2
A priori	2	2.1
Kaiser criterion: Eigenvalue > 1 rule[92]	46	48.4
Cattell scree test[93]	35	33.7
Cattell-Nelson-Gorsuch objective scree[21]	1	1.1
Minimum average partial[108]	1	1.1
Parallel analysis[107]	4	4.2
Minimum proportion of variance accounted for in solution	5	5.3
Minimum number of items per factor	18	18.9
Conceptual interpretability/meaningfulness	21	22.1
Chi-square statistic	3	3.2
Mokken scale analysis[112]	1	1.1
Simple structure	1	1.1
Minimum internal consistency per scale	1	1.1
Not reported	11	11.6
Factor loadings reported		
All factor loadings for all items	33	34.7
Limited loadings	30	31.6
None	31	32.6

* The EFA total indicates those analyses that employed an EFA extraction method (e.g., PAF, maximum likelihood) but does not include the PCAs that authors of reviewed articles incorrectly termed *EFA*.

† For this instance, both PCA and PAF extraction methods were applied; the PAF solution was interpreted.

‡ Orthogonal and oblique totals indicate the number of factor solutions interpreted using each class of rotation methods; subcategories reflect the number of factor rotations applied using each rotation type.

§ Specific rotation types are included under the broader orthogonal and oblique categories.

often are not met in social science data, which limits its applicability.[25] Test–retest reliability and stability are, however, accessible. Although additional planning is required to incorporate these calculations in the research design, most medical education scenarios should provide this opportunity; yet, in this review, most investigators failed to design for this data collection. Approximately 10% of the instruments reviewed did include either multiple raters for an individual or a single rater who rated multiple individuals, but interrater and intrarater reliabilities were not consistently reported.

Researchers reported evidence based on relationships with other variables for few instruments within this review. Specifically, although divergent validity supported roughly 40% of the instruments, most instruments did not have supporting criterion, discriminant, and convergent evidence. This is unfortunate. The relationship between the measure and a theoretically related or unrelated measure, the demonstration of the measure's ability to predict relevant performance, and/or evidence of group differences in scores based on previous theory provide important support for proposed inferences. For example, Lie and colleagues[61] found scores on the Interpreter Scale—an interpreter-led assessment of medical student skills in working with interpreters—correlated with scores on the patient-completed Interpreter Impact Rating Scale and the faculty-completed Faculty Observer Rating Scale; this provides convergent evidence in support of the instrument. Further, Haidet and colleagues[47] examined both concurrent criterion and discriminant validity evidence of the CONNECT instrument through testing of hypothesized relationships between subscale scores and previously validated instruments. In general, evidence based on relationships with other variables is only as strong as the reliability and validity of the associated variables. Therefore, for the instruments reviewed, perhaps the researchers did not identify in the literature rigorously tested measures to apply to investigate validity based on relationships with other variables.

It should be noted that almost all instruments included in this review

were new or revised from an original version. This implies that the first step in establishing evidence for validity would include work on the new or revised instrument's content, structure, and relationship to the theoretical foundation. It is possible that the authors of the studies reviewed are conducting further research with these instruments to provide additional evidence; however, this cannot be commented on given the available evidence. What can be reiterated is the importance of pursuing validity evidence from each source to the extent possible and working to develop a body of literature that uses an instrument across relevant samples and contexts to help improve medical educators' or researchers' confidence in the conclusions they draw from these measures.

Factor analysis

Factor analysis is a large-sample procedure, yet just 25% of analyses in this review met the recommended minimum sample size of 300 participants.[18,24] Larger sample sizes generally produce more stable factor structures and better approximate population parameters. As an alternative metric to absolute sample sizes, participant-to-item ratios from 3:1 to 10:1 are considered best practice.[21,96–98] Although absolute sample size recommendations were not met, more than 50% of analyses in this review met or exceeded the 10:1 recommended participant-to-item ratio.

In selecting factor analysis methods to apply to the sample data, PCA was the predominant model of analysis and extraction method used in the reviewed analyses, despite clear statements in the literature that PCA is not appropriate for instrument development.[20,21,96,99–105] Only 17% of the studies appropriately employed EFA, as determined by this review. Some authors misused terminology and reported that they conducted EFA when they actually used PCA. These two models are not interchangeable: PCA tends to inflate factor loadings, underestimate correlations between factors, and retain error in the model, limiting the potential for the factor structure to be replicated in other samples or confirmed through CFA. Further, when data quality is poor, PCA and EFA may lead to distinctly different results (e.g., different subscale and total scores on an assessment) that can affect the application of an instrument in research and practice.[96,104,105]

Within the exploratory factor model, selection of a rotation method should derive from theoretical or empirical evidence that may suggest correlations, or the lack thereof, between factors. General guidance in the social sciences literature suggests that an oblique rotation is preferred to an orthogonal rotation at first, based on the assumed correlations within sociopsychological constructs.[5,23,24] If evidence suggests that factors are unrelated, an orthogonal rotation may be interpreted instead. Findings from this review indicate that researchers most often applied orthogonal rotations, specifically varimax rotations. Roughly 20% of the analyses included use of oblique rotations. Only about 25% of the analyses reported evidence-based justification for the selected rotation method. Further, some analyses employed orthogonal rotations despite evidence to suggest correlations between factors; this can lead to inflated factor loadings that may influence the interpreted solution and subsequent score calculations.

For nearly 50% of the factor analyses, investigators used only a single criterion to determine the number of factors to retain from the rotated solution. They most often employed the Cattell scree test[93] and Kaiser eigenvalue greater than one rule,[92] though the latter has been largely discredited as the least accurate criterion.[19,20,23,106] Both of these methods tend to overestimate the number of factors to retain, particularly as the number of variables increases. Only a handful of studies made use of more robust, accurate options: for example, parallel analysis[107] (i.e., generating a random data set and corresponding scree plot using the same number of participants and variables as the real data set and retaining no real data factors that explain less variance than the factors from the random data) or minimum average partial[108] (i.e., extracting factors until all common variance is represented in the extracted factors and only unique variance remains in the matrix). These tools are not included in most statistical software packages and, therefore, are not readily available to most researchers.

Once select factors are retained using multiple recommended criteria, all factor loadings for all items must be reported to best interpret and potentially replicate the factor solution. However, more than 33% of the reviewed analyses failed to provide these complete data, instead reporting loadings only for items that were retained in the factor solution. Further, about 33% of the analyses reported none of the factor loadings; most often, this occurred when the items were not included in the article reviewed. Without this information, it is difficult for the reader to understand which items belong to which factor, how to handle items that did not load on a factor in future administrations of the instrument, and how to calculate subscale scores—essentially, future application of the instrument is limited.

Limitations

The findings and conclusions from this study are tempered by the limitations of this review. The *Standards of Educational and Psychological Testing*[17] provided the framework for the review of reliability and validity evidence. Although this contemporary perspective should drive medical education instrument development, it is evident in previous literature[10,16,25,29] and in this review that traditional validity terminology remains predominant in the medical education literature. Some efforts have been made to communicate the contemporary perspective to medical education researchers and practitioners,[9,12–14,25,29,94,109–111] yet their exposure to these concepts may be limited, which may influence the scope of techniques for establishing validity evidence identified in this review. Further, this review's eligibility criteria limited its scope to instrument development articles that specifically employed EFA. Because EFA is a technique most appropriate in the early developmental stages of a new or revised instrument, researchers may be unlikely to engage in longitudinal analysis or further data collection that would allow for investigation of some sources of validity evidence.

Conclusions

In conclusion, what seems to be lacking in current medical education instrument development practice is evidence to indicate how scores on the instrument

relate to other theoretically related or unrelated variables, how scores may predict important expected outcomes, or whether scores remain stable or change over time as anticipated by the theoretical understanding of the construct. Investigation of these sources of evidence, which are critical to the development of a well-rounded argument for the reliability and validity of an instrument, requires resources and more complex research designs, including longitudinal designs. Moving forward, researchers are encouraged to build bodies of research around these and other measurements. Further, this review's findings suggest that the evidence available to support construct validity based on internal structure often rests on inappropriate factor analysis methodology (when methodology is reported). Yet, medical educators and other readers may not be expected to understand the complexities of factor analysis. This point, coupled with these findings, highlights the need for development of additional expertise within the medical education research community and a peer-review process that selects for sound methodological techniques. Researchers and educators should be cautious in adopting and applying instruments from the literature without carefully considering the available supporting evidence. Peer reviewers should be asked to promote instrument development research more consistent with best practices. Aligning current practices in factor analysis and other techniques for establishing validity evidence with best practices can improve instrumentation and lead to better informed inferences about learners and programs across the continuum of medical education.

Acknowledgment: The author sincerely thanks Kelly Lockeman for the time and interest she invested in this study as a coder and contributor to the development of the data abstraction form and coding manual.

Funding/Support: Funding for this project was provided by Pfizer Medical Education Group (grant number 035168).

Other disclosures: None.

Ethical approval: Not applicable.

References

1 Moore DE Jr, Green JS, Gallis HA. Achieving desired results and improved outcomes: Integrating planning and assessment throughout learning activities. J Contin Educ Health Prof. 2009;29:1–15.
2 Henson RK, Roberts JK. Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. Educ Psychol Meas. 2006;66:393–416.
3 Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psych Methods. 2006;4:272–299.
4 Henson RK, Capraro RM, Capraro MM. Reporting practices and use of exploratory factor analyses in educational research journals: Errors and explanation. Res Sch. 2004;11:61–72.
5 Norris M, Lecavalier L. Evaluating the use of exploratory factor analysis in developmental disability psychological research. J Autism Dev Disord. 2010;40:8–20.
6 Park HS, Dailey R, Lemus D. The use of exploratory factor analysis and principal components analysis in communication research. Hum Commun Res. 2002;28: 562–577.
7 Pohlmann JT. Use and interpretation of factor analysis in the Journal of Educational Research. J Educ Res. 2004;98:14–22.
8 Worthington RL, Whittaker TA. Scale development research: A content analysis and recommendations for best practice. Couns Psychol. 2006;34:806–838.
9 Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. J Gen Intern Med. 2004;19:971–977.
10 Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. Med Educ. 2002;36:73–91.
11 Jha V, Bekker HL, Duffy SR, Roberts TE. A systematic review of studies assessing and facilitating attitudes towards professionalism in medicine. Med Educ. 2007;41:822–829.
12 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: A review of published validity evidence. Med Educ. 2011;45:329–338.
13 Ratanawongsa N, Thomas PA, Marinopoulos SS, et al. The reported validity and reliability of methods for evaluating continuing medical education: A systematic review. Acad Med. 2008;83:274–283.
14 Shaneyfelt T, Baum KD, Bell D, et al. Instruments for evaluating education in evidence-based practice: A systematic review. JAMA. 2006;296:1116–1127.
15 Tian J, Atkinson NL, Portnoy B, Gold RS. A systematic review of evaluation in formal continuing medical education. J Contin Educ Health Prof. 2007;27:16–27.
16 Veloski JJ, Fields SK, Boex JR, Blank LL. Measuring professionalism: A review of studies with instruments reported in the literature between 1982 and 2002. Acad Med. 2005;80:366–370.
17 American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 1999.
18 Comrey AL, Lee HB. A First Course in Factor Analysis. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1992.
19 DeVellis RF. Scale Development: Theory and Applications. 2nd ed. Thousand Oaks, Calif: Sage Publications; 2003.
20 Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assess. 1995;7:286–299.
21 Gorsuch RL. Factor Analysis. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum; 1983.
22 Mulaik SA. Foundations of Factor Analysis. 2nd ed. Boca Raton, Fla: CRC Press; 2009.
23 Reise SP, Waller NG, Comrey AL. Factor analysis and scale revision. Psychol Assess. 2000;12:287–297.
24 Tabachnick BG, Fidell LS. Using Multivariate Statistics. 5th ed. Boston, Mass: Pearson; 2007.
25 Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. New York, NY: Oxford University Press; 2008.
26 Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52:281–302.
27 Messick S. The standard program: Meaning and values in measurement and evaluation. Am Psychol. 1975;30:955–956.
28 Messick S. Test validity and the ethics of assessment. Am Psychol. 1980;35:1012–1027.
29 Artino AR Jr, Durning SJ, Creel AH. AM Last Page. Reliability and validity in educational measurement. Acad Med. 2010;85:1545.
30 Aramesh K, Mohebbi M, Jessri M, Sanagou M. Measuring professionalism in residency training programs in Iran. Med Teach. 2009;31:e356–e361.
31 Aukes LC, Geertsma J, Cohen-Schotanus J, Zwierstra RP, Slaets JP. The development of a scale to measure personal reflection in medical practice and education. Med Teach. 2007;29:177–182.
32 Boor K, Scheele F, van der Vleuten CP, Scherpbier AJ, Teunissen PW, Sijtsma K. Psychometric properties of an instrument to measure the clinical learning environment. Med Educ. 2007;41:92–99.
33 Campbell C, Lockyer J, Laidlaw T, Macleod H. Assessment of a matched-pair instrument to examine doctor–patient communication skills in practising doctors. Med Educ. 2007;41:123–129.
34 Carruthers S, Lawton R, Sandars J, Howe A, Perry M. Attitudes to patient safety amongst medical students and tutors: Developing a reliable and valid measure. Med Teach. 2009;31:e370–e376.
35 Chou B, Bowen CW, Handa VL. Evaluating the competency of gynecology residents in the operating room: Validation of a new assessment tool. Am J Obstet Gynecol. 2008;199:571.e1–571.e5.
36 Colletti JE, Flottemesch TJ, O'Connell TA, Ankel FK, Asplin BR. Developing a standardized faculty evaluation in an emergency medicine residency. J Emerg Med. 2010;39:662–668.
37 Cruess R, McIlroy JH, Cruess S, Ginsburg S, Steinert Y. The Professionalism Mini-evaluation Exercise: A preliminary investigation. Acad Med. 2006;81(10 suppl):S74–S78.

38 Di Lillo M, Cicchetti A, Lo Scalzo A, Taroni F, Hojat M. The Jefferson Scale of Physician Empathy: Preliminary psychometrics and group comparisons in Italian physicians. Acad Med. 2009;84:1198–1202.
39 Dimoliatis ID, Vasilaki E, Anastassopoulos P, Ioannidis JP, Roff S. Validation of the Greek translation of the Dundee Ready Education Environment Measure (DREEM). Educ Health (Abingdon). 2010;23:348.
40 Donnon T, Woloschuk W, Myhre D. Issues related to medical students' engagement in integrated rural placements: An exploratory factor analysis. Can J Rural Med. 2009;14:105–110.
41 El-Zubeir M, Rizk DE, Al-Khalil RK. Are senior UAE medical and nursing students ready for interprofessional learning? Validating the RIPL scale in a Middle Eastern context. J Interprof Care. 2006;20:619–632.
42 Eslaminejad T, Masood M, Ngah NA. Assessment of instructors' readiness for implementing e-learning in continuing medical education in Iran. Med Teach. 2010;32:e407–e412.
43 Flin R, Patey R, Jackson J, Mearns K, Dissanayaka U. Year 1 medical undergraduates' knowledge of and attitudes to medical error. Med Educ. 2009;43:1147–1155.
44 Frye AW, Sierpina VS, Boisaubin EV, Bulik RJ. Measuring what medical students think about complementary and alternative medicine (CAM): A pilot study of the complementary and alternative medicine survey. Adv Health Sci Educ Theory Pract. 2006;11:19–32.
45 Gaspar MF, Pinto AM, da Conceicao HCF, da Silva JAP. A questionnaire for listening to students' voices in the assessment of teaching quality in a classical medical school. Assess Eval High Educ. 2008;33:445–453.
46 Gooneratne IK, Munasinghe SR, Siriwardena C, Olupeliyawa AM, Karunathilake I. Assessment of psychometric properties of a modified PHEEM questionnaire. Ann Acad Med Singap. 2008;37:993–997.
47 Haidet P, O'Malley KJ, Sharf BF, Gladney AP, Greisinger AJ, Street RL Jr. Characterizing explanatory models of illness in healthcare: Development and validation of the CONNECT instrument. Patient Educ Couns. 2008;73:232–239.
48 Harlak H, Dereboy C, Gemalmaz A. Validation of a Turkish translation of the Communication Skills Attitude Scale with Turkish medical students. Educ Health (Abingdon). 2008;21:55.
49 Harvey BJ, Rothman AI, Frecker RC. A confirmatory factor analysis of the Oddi Continuing Learning Inventory (OCLI). Adult Educ. 2006;56:188–200.
50 Helayel PE, da Conceição DB, da Conceição MJ, Boos GL, de Toledo GB, de Oliveira Filho GR. The attitude of anesthesiologists and anesthesiology residents of the CET/SBA regarding upper and lower limb nerve blocks. Rev Bras Anestesiol. 2009;59:332–340.
51 Hendry GD, Ginns P. Readiness for self-directed learning: Validation of a new scale with medical students. Med Teach. 2009;31:918–920.
52 Hojat M, Veloski J, Nasca TJ, Erdmann JB, Gonnella JS. Assessing physicians' orientation toward lifelong learning. J Gen Intern Med. 2006;21:931–936.
53 Hojat M, Veloski JJ, Gonnella JS. Measurement and correlates of physicians' lifelong learning. Acad Med. 2009;84:1066–1074.
54 Holt KD, Miller RS, Philibert I, Heard JK, Nasca TJ. Residents' perspectives on the learning environment: Data from the Accreditation Council for Graduate Medical Education resident survey. Acad Med. 2010;85:512–518.
55 Kane GC, Gotto JL, Mangione S, West S, Hojat M. Jefferson Scale of Patient's Perceptions of Physician Empathy: Preliminary psychometric data. Croat Med J. 2007;48:81–86.
56 Kataoka HU, Koide N, Ochi K, Hojat M, Gonnella JS. Measurement of empathy among Japanese medical students: Psychometrics and score differences by gender and level of medical education. Acad Med. 2009;84:1192–1197.
57 Klein B, McCall L, Austin D, Piterman L. A psychometric evaluation of the Learning Styles Questionnaire: 40-item version. Br J Educ Technol. 2007;38:23–32.
58 Lam TP, Wong JG, Ip MS, Lam KF, Pang SL. Psychological well-being of interns in Hong Kong: What causes them stress and what helps them. Med Teach. 2010;32:e120–e126.
59 Leenstra JL, Beckman TJ, Reed DA, et al. Validation of a method for assessing resident physicians' quality improvement proposals. J Gen Intern Med. 2007;22:1330–1334.
60 Leung KK, Wang WD. Validation of the Tutotest in a hybrid problem-based learning curriculum. Adv Health Sci Educ Theory Pract. 2008;13:469–477.
61 Lie D, Bereknyei S, Braddock CH 3rd, Encinas J, Ahearn S, Boker JR. Assessing medical students' skills in working with interpreters during patient encounters: A validation study of the Interpreter Scale. Acad Med. 2009;84:643–650.
62 Lin GA, Beck DC, Stewart AL, Garbutt JM. Resident perceptions of the impact of work hour limitations. J Gen Intern Med. 2007;22:969–975.
63 Lockyer JM, Violato C, Fidler H, Alakija P. The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. Arch Pathol Lab Med. 2009;133:1301–1308.
64 McCormack WT, Lazarus C, Stern D, Small PA Jr. Peer nomination: A tool for identifying medical student exemplars in clinical competence and caring, evaluated at three medical schools. Acad Med. 2007;82:1033–1039.
65 McLaughlin K, Vitale G, Coderre S, Violato C, Wright B. Clerkship evaluation—What are we measuring? Med Teach. 2009;31:e36–e39.
66 McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ. 2006;6:42.
67 Menachery EP, Knight AM, Kolodner K, Wright SM. Physician characteristics associated with proficiency in feedback skills. J Gen Intern Med. 2006;21:440–446.
68 Mihalynuk TV, Coombs JB, Rosenfeld ME, Scott CS, Knopp RH. Survey correlations: Proficiency and adequacy of nutrition training of medical students. J Am Coll Nutr. 2008;27:59–64.
69 Mitchell R, Regan-Smith M, Fisher MA, Knox I, Lambert DR. A new measure of the cognitive, metacognitive, and experiential aspects of residents' learning. Acad Med. 2009;84:918–926.
70 Nagraj S, Wall D, Jones E. The development and validation of the mini-surgical theatre educational environment measure. Med Teach. 2007;29:e192–e197.
71 Orlander JD, Wipf JE, Lew RA. Development of a tool to assess the team leadership skills of medical residents. Med Educ Online. 2006;11:1–6.
72 Park ER, Chun MB, Betancourt JR, Green AR, Weissman JS. Measuring residents' perceived preparedness and skillfulness to deliver cross-cultural care. J Gen Intern Med. 2009;24:1053–1056.
73 Pentzek M, Abholz HH, Ostapczuk M, Altiner A, Wollny A, Fuchs A. Dementia knowledge among general practitioners: First results and psychometric properties of a new instrument. Int Psychogeriatr. 2009;21:1105–1115.
74 Reinders ME, Blankenstein AH, Knol DL, de Vet HC, van Marwijk HW. Validity aspects of the patient feedback questionnaire on consultation skills (PFC), a promising learning instrument in medical education. Patient Educ Couns. 2009;76:202–206.
75 Riquelme A, Herrera C, Aranis C, Oporto J, Padilla O. Psychometric analyses and internal consistency of the PHEEM questionnaire to measure the clinical learning environment in the clerkship of a medical school in Chile. Med Teach. 2009;31:e221–e225.
76 Rogers ME, Creed PA, Searle J. The development and validation of social cognitive career theory instruments to measure choice of medical specialty and practice location. J Career Assess. 2009;17:324–337.
77 Roh MS, Hahm BJ, Lee DH, Suh DH. Evaluation of empathy among Korean medical students: A cross-sectional study using the Korean Version of the Jefferson Scale of Physician Empathy. Teach Learn Med. 2010;22:167–171.
78 Sargeant J, Hill T, Breau L. Development and testing of a scale to assess interprofessional education (IPE) facilitation skills. J Contin Educ Health Prof. 2010;30:126–131.
79 Short LM, Alpert E, Harris JM Jr, Surprenant ZJ. A tool for measuring physician readiness to manage intimate partner violence. Am J Prev Med. 2006;30:173–180.
80 Singer Y, Carmel S. Teaching end-of-life care to family medicine residents—What do they learn? Med Teach. 2009;31:e47–e50.
81 Sladek RM, Phillips PA, Bond MJ. Measurement properties of the Inventory of Cognitive Bias in Medicine (ICBM). BMC Med Inform Decis Mak. 2008;8:20.
82 Sodano SM, Richard GV. Construct validity of the medical specialty preference inventory: A critical analysis. J Vocat Behav. 2009;74:30–37.

83 Tian J, Atkinson NL, Portnoy B, Lowitt NR. The development of a theory-based instrument to evaluate the effectiveness of continuing medical education. Acad Med. 2010;85:1518–1525.

84 Tromp F, Vernooij-Dassen M, Kramer A, Grol R, Bottema B. Behavioural elements of professionalism: Assessment of a fundamental concept in medical care. Med Teach. 2010;32:e161–e169.

85 Tsai TC, Lin CH, Harasym PH, Violato C. Students' perception on medical professionalism: The psychometric perspective. Med Teach. 2007;29:128–134.

86 Vasan NS, DeFouw DO, Compton S. A survey of student perceptions of team-based learning in anatomy curriculum: Favorable views unrelated to grades. Anat Sci Educ. 2009;2:150–155.

87 Vieira JE. The postgraduate hospital educational environment measure (PHEEM) questionnaire identifies quality of instruction as a key factor predicting academic achievement. Clinics (Sao Paulo). 2008;63:741–746.

88 Wall D, Clapham M, Riquelme A, et al. Is PHEEM a multi-dimensional instrument? An international perspective. Med Teach. 2009;31:e521–e527.

89 Wetzel AP, Mazmanian PE, Hojat M, et al. Measuring medical students' orientation toward lifelong learning: A psychometric evaluation. Acad Med. 2010;85(10 suppl):S41–S44.

90 Wittich CM, Beckman TJ, Drefahl MM, et al. Validation of a method to measure resident doctors' reflections on quality improvement. Med Educ. 2010;44:248–255.

91 Wright SM, Levine RB, Beasley B, et al. Personal growth and its correlates during residency training. Med Educ. 2006;40:737–745.

92 Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas. 1960;20:141–151.

93 Cattell RB. The scree test or the number of factors. Multivariate Behav Res. 1966;1:245–276.

94 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med. 2006;119:166.e7–166.16.

95 McDonald RP. Test Theory: A Unified Treatment. Mahwah, NJ: Lawrence Erlbaum Associates; 1999.

96 Costello AB, Osborne J. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assess Res Eval. 2005;10:1–9.

97 Everitt BS. Multivariate analysis: The need for data and other problems. Br J Pscyhiatry. 1975;126:237–240.

98 Tinsley HEA, Tinsley DJ. Uses of factor analysis in counseling psychology research. J Couns Psychol. 1987;34:414–424.

99 Bentler PM, Kano Y. On the equivalence of factors and components. Multivariate Behav Res. 1990;25:67–74.

100 Gorsuch R. Common factor analysis versus component analysis: Some well and little known facts. Multivariate Behav Res. 1990;25:33.

101 Mulaik SA. Blurring the distinctions between component analysis and common factor analysis. Multivariate Behav Res. 1990;25:53–59.

102 Snook SC, Gorsuch RL. Component analysis versus common factor analysis: A Monte-Carlo study. Psychol Bull. 1989;106:148–154.

103 Widaman KF. Bias in pattern loadings represented by common factor analysis and component analysis. Multivariate Behav Res. 1990;25:89–95.

104 Widaman KF. Common factor analysis versus principal component analysis: Differential bias in representing model parameters? Multivariate Behav Res. 1993;28:263.

105 Widaman KF. Common factors versus components: Principals and principles, errors and misconceptions. In: Cudeck J, MacCallum RC, eds. Factor Analysis at 100: Historical Developments and Future Directions. Mahwah, NJ: Lawrence Erlbaum; 2007.

106 Zwick W, Velicer W. Comparison of five rules for determining the number of components to retain. Psychol Bull. 1986;99:432–442.

107 Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;30:179–185.

108 Bandalos DL, Boehm-Kaufman MR. Four common misconceptions in exploratory factor analysis. In: Lance CE, Vandenberg RJ, eds. Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in the Organizational and Social Sciences. New York, NY: Routledge Taylor & Francis Group; 2009.

109 Andreatta PB, Gruppen LD. Conceptualising and classifying validity evidence for simulation. Med Educ. 2009;43:1028–1035.

110 Downing SM. Validity: On meaningful interpretation of assessment data. Med Educ. 2003;37:830–837.

111 Downing SM. Reliability: On the reproducibility of assessment data. Med Educ. 2004;38:1006–1012.

Reference in Table 3 Only

112 Mokken RJ. A Theory and Procedure of Scale Analysis With Applications in Political Research. New York, NY: De Gruyter; 1971.