2011

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

Angela Wetzel
Virginia Commonwealth University

FACTOR ANALYSIS METHODS AND VALIDITY EVIDENCE:

A SYSTEMATIC REVIEW OF INSTRUMENT DEVELOPMENT ACROSS THE

CONTINUUM OF MEDICAL EDUCATION

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at Virginia Commonwealth University

by

Angela Payne Wetzel
Bachelor of Arts, University of Virginia, 2003
Master of Education, Virginia Commonwealth University, 2005

Director: James H. McMillan, Ph.D.
Professor, Foundations of Education
School of Education

Virginia Commonwealth University
Richmond, VA
April 2011

Acknowledgment

With the following dissertation as the final bookend to my doctoral education, I am overwhelmed with gratitude to those who have helped to make this achievement a reality. To Denny Hoban, who sparked my initial curiosity in research and evaluation, encouraged and counseled me as I pursued this graduate degree, and who cared for me as his own daughter, I miss you, and I wish you could share this moment with me. I am grateful to all my graduate faculty members in the program who shared immeasurable knowledge and enthusiasm for the content; it is infectious. I want to thank Paul Mazmanian for creating opportunities to keep my feet grounded in medical education and for providing unending support not only of my work, but also for me personally as I juggled school, this dissertation, work, and my family. To my chair and advisor, Jim McMillan, thank you for encouraging me throughout the program, for reminding me of the priorities, and for caring about me not only as a developing scholar but as a person. Thank you so much for your support of my dissertation topic, though a less traditional methodology, it has been incredibly valuable for me. And thank you to all of my committee members, Lisa Abrams, Teresa Carter, and Levent Dumeni, for your time, your thoughtful comments, and encouraging words. This dissertation is a reflection not only of my individual work but of your time and interest; thank you for helping me do better work. I would be remiss not to thank Pfizer Medical Education Group for their financial support of this project.

Last, but certainly not least, I must thank my family. To my in-laws, Jim and Mary Lee, who watched infant Caroline so I would not miss class and who continued to check in on my progress over the years, thank you. Often, this experience can be isolating; it is so nice to have others check in on your progress and well-being. To Mom and Dad, I owe so much. You instilled in me early the value of hard work and the importance of education, and you always encouraged me to be my best. Thank you for providing a solid foundation of love and support that has maintained me my whole life. My dear Caroline, I do this work to better myself, to better our family, and to be a better mother and role model for you. I hope you grow up to understand that you can do and be anything with focused motivation and a love of learning, knowing I will always be behind you 100%. And to Andy, my husband and very best friend, you have believed in me and supported me more than anyone. I am a better person, in all ways, because you are in my life, and I thank you a thousand times over for your love, your patience, and your encouragement. I am blessed that you have been by my side through this journey.

Table of Contents

Tables

Figures

Abstract

FACTOR ANALYSIS METHODS AND VALIDITY EVIDENCE: A SYSTEMATIC REVIEW OF INSTRUMENT DEVELOPMENT ACROSS THE CONTINUUM OF MEDICAL EDUCATION

By Angela Payne Wetzel, Ph.D.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University, 2011

Director: James H. McMillan, Ph.D.
Professor, Foundations of Education
School of Education

Previous systematic reviews indicate a lack of reporting of reliability and validity evidence in subsets of the medical education literature. Psychology and general education reviews of factor analysis also indicate gaps between current and best practices; yet, a comprehensive review of exploratory factor analysis in instrument development across the continuum of medical education had not been previously identified. Therefore, the purpose for this study was critical review of instrument development articles employing exploratory factor or principal component analysis published in medical education (2006-2010) to describe and assess the reporting of

methods and validity evidence based on the *Standards for Educational and Psychological Testing* and factor analysis best practices.

Data extraction of 64 articles measuring a variety of constructs that have been published throughout the peer-reviewed medical education literature indicate significant errors in the translation of exploratory factor analysis best practices to current practice. Further, techniques for establishing validity evidence tend to derive from a limited scope of methods including reliability statistics to support internal structure and support for test content. Instruments reviewed for this study lacked supporting evidence based on relationships with other variables and response process, and evidence based on consequences of testing was not evident.

Findings suggest a need for further professional development within the medical education researcher community related to 1) appropriate factor analysis methodology and reporting and 2) the importance of pursuing multiple sources of reliability and validity evidence to construct a well-supported argument for the inferences made from the instrument. Medical education researchers and educators should be cautious in adopting instruments from the literature and carefully review available evidence. Finally, editors and reviewers are encouraged to recognize this gap in best practices and subsequently to promote instrument development research that is more consistent through the peer-review process.

Chapter 1

Introduction

Background for the Study

Measurement is a core element of science. Some disciplines, particularly physical sciences, concentrate on the measurement of variables that can be directly observed and thus measured. Whereas, across the social sciences including education, researchers often investigate phenomena that cannot be directly observed and measured. Proxy measures, traditionally in the form of tests or questionnaires, are often developed to enable measurement of these underlying constructs (DeVellis, 2003). If prudent instrument development is practiced, quality instrumentation that serves as an accurate and precise measure of the construct of interest can be created. However, application of measurement in research and practice in the absence of rigorous instrument development can lead to erroneous conclusions.

Medical education, compared to general education or more broadly the social sciences, is not unique in the need for measurement. Across the medical education continuum, including undergraduate, graduate, and continuing medical education, written examinations, questionnaires, performance based checklists, objective structured clinical examinations, and standardized patient examinations are measurement tools frequently used for assessment and evaluation of outcomes ranging from the individual learner level to the patient and community health level (Moore, Green, & Gallis, 2009). Thus, quality measurement is critical in medical education.

The medical education continuum is made up of three stages: (a) undergraduate medical education, (b) graduate medical education and (c) continuing medical education,

with each stage representing a component of the longitudinal training and professional development of physicians (See Figure 1). Undergraduate medical education (UME) refers to the first four years of medical training leading to the doctorate of medicine (M.D.) degree. Currently, 133 medical schools in the United States are accredited by the Liaison Committee on Medical Education (LCME) to award the M.D. degree. Education at the undergraduate level focuses on fundamentals of medical knowledge, clinical skills, and limited, supervised practice of medicine in hospital and ambulatory settings. Once a student graduates from an LCME-accredited medical school, he or she becomes eligible to apply for a residency position with a graduate medical education (GME) program accredited by the Accreditation Council for Graduate Medical Education (ACGME). Where UME focuses on broad medical knowledge and basic skills, GME provides in-depth knowledge and skills training in a specialty area of medicine (e.g., Internal Medicine, Obstetrics and Gynecology, or Psychiatry). The graduate medical education phase, or residency, may be three to seven years in duration, though most last four or five depending on the chosen specialty. Resident physicians practice medicine under the supervision of fully licensed physicians. Successful completion of the residency program and specialty board certification examinations is required to practice medicine independently. Across the undergraduate and graduate training years, students will sit for three written and one clinical United States Medical Licensing Examinations (USMLE); passing scores on all four exams are required to receive a medical license. Once in practice, physicians are mandated to participate in continuing medical education (CME) through programs accredited by the Accreditation Council for Continuing Medical Education (ACCME) to maintain licensure and certification.

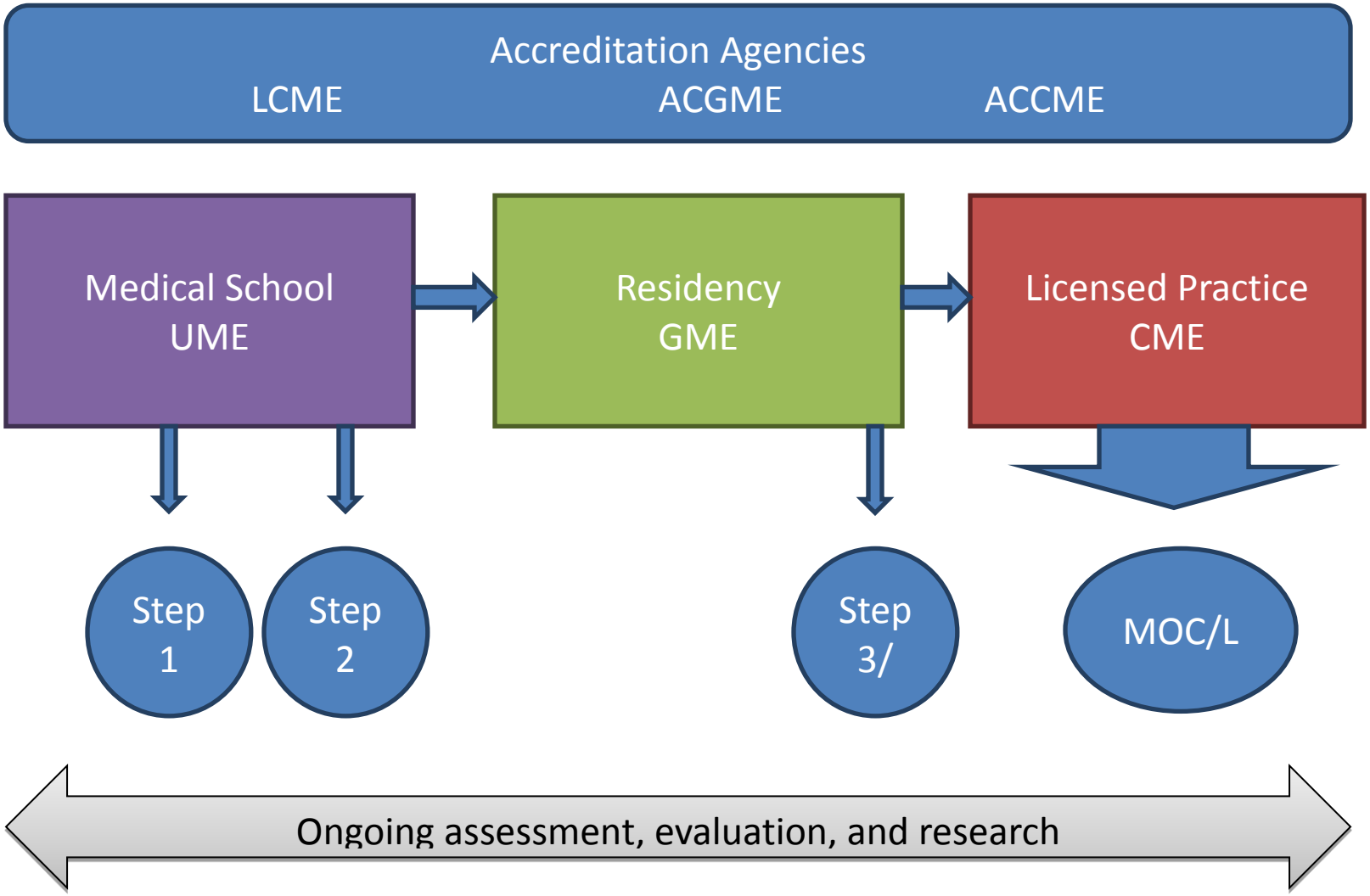

Figure 1. Overview of the medical education continuum

Apart from national standardized and USMLE examinations authored by the National Board of Medical Examiners (NBME), most assessment and evaluation instruments in medical education are developed at the institutional level, often with little to no funding, by medical educators with varying expertise in measurement and research (Carline, 2004; Reed, Cook, Beckman, Levine, Kern, & Wright, 2007; Reed, Kern, Levine, & Wright, 2005; Shea, Arnold, & Mann, 2004). Most of the data from these instruments are used in formative and summative ways for assessing students and evaluating programs and are incorporated into medical education research endeavors. Overall, medical education research has been asked to adopt into practice established research methodological standards to promote robust research for the field (Albert & Reeves, 2010; Andreatta & Gruppen, 2009; Cook & Beckman, 2006; Downing, 2004; Downing, 2003; Schonrock-Adema, Heijne-Penninga, van Hell, & Cohen-Schotanus, 2009). Specifically, efforts continue to be made to communicate best practices for instrument development, validation, and reporting throughout the medical education research practitioner community (Andreatta & Gruppen, 2009; Boulet, De Champlain, & McKinley, 2003; Cook & Beckman, 2006; Downing, 2003; Downing, 2004; Schonrock-Adema et al., 2009; Streiner & Norman, 2008); yet, how extensively best practices have been implemented remains unclear. Therefore, it is necessary to gain a better understanding of current practice to inform the work of medical education researchers and medical educators who make critical decisions regarding quality instruments for application in their programs.

Overview of the Literature

Instrument Development. Within the social sciences, psychometrics emerged as the field of study underlying the theory and techniques of educational and psychological measurement. Initially, the field developed from an interest in ability testing and then expanded into measurement of other social or psychological latent variables. Latent variable is a term used to refer to the construct or phenomenon of interest that cannot be directly observed or measured. Much of the current work with psychometrics involves the development and testing of instruments including assessments and questionnaires to accurately define and quantify latent variables (DeVellis, 2003).

Although more than one sequence of steps for instrument development has been proposed, a common set of practices can be identified among authors' recommendations: (a) clearly define what is to be measured, (b) generate an item pool, (c) ask experts to review the item pool, (d) format and pilot test the items with a sample from the target population, (e) theoretically and empirically evaluate the items, and (f) revise items and establish optimal scale length (American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), 1999; DeVellis, 2003; Streiner & Norman, 2008). In rigorous instrument development and psychometric testing, each step mentioned previously generates sources of evidence to support the validity of inferences made from the test scores. This supporting evidence for validity should be reported in instrument development literature to allow the consumer of the instrument to capably appraise its fit for assessment or evaluation needs based on how the construct is defined, the nature of the target population, the psychometrics, and other key characteristics.

Sources of Validity Evidence. The *Standards for Educational and Psychological Testing* (AERA et al., 1999) provides preeminent guidance on sources of validity evidence. Under the contemporary conceptualization purported in the *Standards* (AERA et al., 1999), validity is a unitary concept established through the presentation of accumulated evidence based on test content, response processes, internal structure, relationships with other variables, and consequences of testing. Although over a decade old, this new understanding of validity has been somewhat slow to replace the traditional concept of multiple types of validity (e.g., face validity, content validity, or discriminant validity). However, the *Standards* (AERA et al., 1999) are the leading source for conceptualizing validity evidence, and support for this framework is evident in medical education literature including calls for improved practice and recent reviews of validity and reliability evidence (Andreatta & Gruppen, 2009; Beckman, Ghosh, Cook, Erwin, & Mandrekar, 2004; Cook & Beckman, 2006; Ratanawongsa, Thomas, Marinopoulos, Dorman, Wilson, Ashar, Magaziner, Miller, Prokopowicz, Qayyum, & Bass, 2008; Shaneyfelt, Baum, Bell, Feldstein, Houston, Kaatz, Whelan, & Green, 2006; Veloski, Fields, Boex, & Blank, 2005). Yet, currently, no comprehensive review of the medical education literature for the use of techniques for establishing validity in instrument development has been identified.

Factor Analysis. Exploratory factor analysis is often applied in medical education research; it is one of the most useful methods in instrument development for establishing validity evidence based on internal structure (Henson & Roberts, 2006; Kieffer, 1999). Methodological decisions and justification for these decisions should be based on best practices and clearly reported in the literature; otherwise, the potential for

verification or replication by other researchers is limited (Henson & Roberts, 2006; Pohlmann, 2004). Yet, the complexity of factor analytic techniques can make effective utilization of the procedure challenging (Floyd & Widamen, 1995). Although factor analysis best practices lack endorsement by a single, authoritative source, a framework for best practices based on a common set of critical methodological decisions and reporting requirements can be developed from the literature. Clear reporting and justification for sample size criteria, model of analysis, criteria for selection of extraction and rotation methods, and criteria for factor retention is essential (Comrey & Lee, 1992; Floyd & Widaman, 1995; Gorsuch, 1983; Reise, Waller, & Comrey, 2000; Tabachnick & Fidell, 2007). What remains unclear in medical education research is the extent to which best practices associated with factor analysis have been implemented in instrument development.

Reviews of Validity Evidence. A number of previous reviews evaluated the reporting of reliability and validity evidence in medical education literature (Beckman et al., 2004; Hutchinson, Aitken, & Hayes, 2002; Jha, Bekker, Duffy, & Roberts, 2007; Lubarsky, Charlin, Cook, Chalk, van der Vleuten, 2011; Ratanawongsa et al., 2008; Shaneyfelt et al., 2006; Veloski et al., 2005). The consensus across findings reflects insufficient reporting of reliability and validity with evidence based on response process, internal structure, and test content most commonly included. Slightly more than half of the reviews of validity evidence were structured around or made reference to the *Standards* (1999) (Beckman et al., 2004; Lubarsky et al., 2011; Ratanawongsa et al., 2008; Shaneyfelt et al., 2006; Veloski et al., 2005). These reviews focused on subsets of instruments (e.g., instruments measuring professionalism (Jha et al., 2007)); yet, a

comprehensive review of reliability and validity evidence in medical education instrument development has not been reported.

Reviews of Factor Analysis. Reviews of factor analysis procedures are published in other fields including psychology and education more generally (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Henson, Capraro, & Capraro, 2004; Henson & Roberts, 2006; Norris & Lecavalier, 2010; Park, Dailey, & Lemus, 2002; Pohlmann, 2004; Worthington & Whittaker, 2006). Results suggest insufficient reporting of methods and results limiting evaluation of the instrument or possible replication. In addition, Henson and Roberts' (2006) findings also illuminate the reliance of researchers on default options in factor analysis statistical software which may not be appropriate for all instruments and research questions. Specifically in medical education, Schonrock-Adema et al. (2009) reported a need for improvement in instrument validation procedures and articulated a short list of necessary steps for effective factor analysis; however, this assessment was based on a limited discussion of educational environment questionnaires not specific to medicine. Further reviews of the literature have not identified a full evaluation of factor analytic techniques in medical education. This proposed systematic review of medical education instrument development aims to fill these two identified gaps, by evaluating factor analysis methods and by reporting validity evidence in medical education instrument development.

Rationale and Purpose for the Study

Clear reporting of instrument development, including evidence for reliability and validity, is a responsibility of the instrument developer; critical evaluation of such evidence is an essential obligation of the instrument consumer. The good faith efforts of

both parties are required for effective instrument development and application. In view of previous research that indicates insufficient reliability and validity evidence in subsets of instrument development literature, and the overall lack of information on factor analysis studies in the field, a comprehensive review of instrument development across the medical education continuum offers a perspective on best practices and opportunities for improvement. These findings should promote better informed instrument development and research, while enabling medical educators to critically select well-developed, validated instruments.

Therefore, the purpose for this study was to critically review instrument development articles employing exploratory factor or principal component analysis published in medical education (2006-2010) to describe and assess the reporting of methods and validity evidence based on the *Standards for Educational and Psychological Testing* (AERA et al., 1999) and factor analysis best practices as they derive from the literature (Comrey & Lee, 1992; Floyd & Widaman, 1995; Gorsuch, 1983; Reise et al., 2000; Tabachnick & Fidell, 2007).

Research Questions

Findings from this study inform the following two research questions.

Within medical education instrument development literature, including undergraduate, graduate, and continuing medical education:

1. To what extent are techniques for establishing validity consistent with the *Standards for Educational and Psychological Testing* (AERA et al., 1999)?

2. To what extent are exploratory factor and principal component analysis methods, data analysis, and reported evidence consistent with factor analytic best practices?

Design and Methods

This research study employed systematic review methodologies. The Cochrane Collaboration is the leader in systematic reviews in healthcare, and the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre) in the United Kingdom is the leader in defining and conducting systematic reviews in the social sciences and public policy. Together, they characterize systematic reviews using three criteria: (a) a comprehensive review of research evidence delimited by eligibility criteria, (b) explicit, transparent, reproducible methods, and (c) a systematic approach to the organization and presentation of findings from the reviewed studies (Evidence for Policy and Practice Information and Co-ordinating Centre, 2010; Green, Higgins, Alderson, Clarke, Mulrow, Oxman, 2008).

Using a search strategy to combine descriptors and keywords related to instrument development (e.g., validity, reliability, measures, factor analysis) with terms delimiting medical education, peer-reviewed articles published 2006 through 2010 were searched through MEDLINE, ERIC, PsycINFO, and CINAHL electronic databases. Reference lists of all included reports were hand searched. Based on a screening of titles, abstracts, and full-text, primary empirical instrument development research articles employing exploratory factor analysis or principal component analysis and published in English were included. Both newly developed and revised instruments were included. Principal components analysis (PCA) studies were included in order to examine how

often PCA was used in place of a common factors model. If a study combines an EFA with a follow up confirmatory factor analysis (CFA), only the EFA methods and reported evidence were reviewed. Studies employing only CFA within instrument development were excluded to narrow the scope of the study for feasibility reasons.

A data extraction form and coding manual, developed from the literature on best practices in instrument development, provided a structure and process for the researcher to systematically extract from each eligible article the factor analysis methods and analysis and reported evidence for establishing validity. This structured data entry form was pilot tested using select peer-reviewed instrument development articles (n = 5) published in 2005, prior to the proposed review time frame of 2006-2010. The pilot test of the data extraction form informed necessary revisions. A second individual with expertise in the content area was trained to use the data extraction form through self-study of the literature on best practices for instrument development and three iterative rounds of coding and review of agreements and disagreements with the researcher. Experience from these three rounds informed further revisions to the form and coding manual. By applying the revised form and manual, the second coder double-coded a randomly selected 10 percent of all reviewed articles. Further, the researcher utilized the revised form and coding manual to code all articles meeting the eligibility criteria for inclusion in the review. Agreement was calculated. Reviewed instruments were categorized by the level of outcome assessed (e.g., level 3A: declarative knowledge, level 3B: procedural knowledge, or level 4: competence) according to the Outcomes Framework accepted in practice in medical education (Moore et al., 2009). Categorization of instruments by outcome offers a meaningful organizational structure to

the results to aid in interpretation. Results present instruments by outcome level and by frequencies and percentages of articles consistent with best practices.

Definition of Terms

Communality: "The proportion of observed variance due to common factors, or the total amount of variance for an item explained by the extracted factors. [Communalities] can range from zero (the variable has no correlation with any other variable in the matrix) to one (the variance of the variable is completely accounted for by the underlying factors). …In PCA, communalities are set to one, as all observed variance is viewed as available to be modeled." (Norris & Lecavalier, 2010, p. 10-11)

Confirmatory factor analysis: CFA is "a much more sophisticated technique [than EFA] used in the advanced stages of the research process to test a theory about latent processes" (Tabachnick & Fidell, 2007, p.609).

Educational outcome: Classification is based on Moore et al. (2009) Outcomes Framework for Assessing Learners and Evaluating Instructional Activities with seven levels: participation, satisfaction, declarative knowledge, procedural knowledge, competence, performance, patient health, community health (See Table 1).

Table 1
Moore et al. (2009) Outcomes Framework

Outcomes Framework	Description
Participation LEVEL 1	Number of learners who participate in the educational activity
Satisfaction LEVEL 2	Degree to which expectations of participants were met regarding the setting and delivery of the educational activity
Learning: Declarative Knowledge LEVEL 3A	The degree to which participants state *what* the educational activity intended them to know
Learning: Procedural Knowledge LEVEL 3B	The degree to which participants state *how* to do what the educational activity intended them to know how to do
Competence LEVEL 4	The degree to which participants *show* in an educational setting *how* to do what the educational activity intended them to be able to do
Performance LEVEL 5	The degree to which participants *do* what the educational activity intended them to be able to do in their practices
Patient health LEVEL 6	The degree to which the health status of patients improves due to changes in the practice behavior of participants
Community health LEVEL 7	The degree to which the health status of a community of patients changes due to changes in the practice behavior of participants

Source: Moore et al. (2009)

Exploratory factor analysis: EFA is performed in the early stages of research "when there is a theory about underlying structure or when the researcher wants to understand underlying structure" (Tabachnick & Fidell, 2007, p.26). "It provides a tool for consolidating variables and for generating hypotheses about underlying processes" (Tabachnick & Fidell, 2007, p.609).

Reliability: "Reliability is concerned with the consistency, stability, and dependability of the scores" from an assessment or questionnaire (McMillan, 2007). Under the conceptualization of validity as a unitary concept, reliability is understood to provide evidence for support of validity based on internal structure and response process (See Table 2).

Validity: Classification is based on the contemporary approach to validity evidence from the *Standards for Educational and Psychological Testing* of the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, The American Psychological Association, and the National Council on Measurement in Education (1999) which considers validity as a unitary concept representing an accumulation of evidence based on five sources: test content, response processes, internal structure, relations to other variables, and consequences of testing. A comparison of the traditional reliability and validity classification system and contemporary framework is presented in Table 2.

Table 2
Comparison of traditional and contemporary approaches to validity evidence

Traditional classification of validity or reliability	Definition	Mapping of traditional to contemporary approach to validity evidence
Construct validity	Degree to which a measure assesses the theoretical construct intended to be measured	"Validity is a unitary concept….All validity is construct validity in this current framework"
Face/content validity	Degree to which an instrument accurately represents the skill or characteristic that it is designed to measure, according to people's experience and available knowledge	Test content validity remains one of five essential sources of evidence, but face validity is no longer considered
Expert review	The use of individuals with expertise in the content area who evaluate the content of the instrument in relation to the defined construct	Test content
Pilot study	A preliminary study conducted with a sample from the target population to determine the clarity and completeness of items and/or initial psychometrics of an instrument	Test content
Test criterion validity: Concurrent evidence	Degree to which an instrument produces the same results as another accepted, validated, or even "gold standard" instrument that measures the same construct	Relationships with other variables
Test criterion validity: Predictive evidence	Degree to which a measure accurately predicts something it should theoretically be able to predict	Relationships with other variables

Convergent evidence	Degree of agreement between measurements of the same construct obtained by different methodologies (e.g., objective versus subjective)	Relationships with other variables
Discriminant evidence	Degree to which a measure produces results different from the results of another measure of a theoretically unrelated construct	Relationships with other variables
Divergent evidence	Ability of a measure to yield different mean values between relevant groups	Relationships with other variables
Intra-rater reliability	Degree to which measurements are the same when repeated by the same person	Response process
Inter-rater reliability	Degree to which measurements are the same when obtained by different people	Response process
Test-retest reliability	Degree to which the same test produces the same results when repeated under the same conditions (around a two week interval)	Response process
Test-retest stability	Degree to which the same test produces the same results when repeated under the same conditions (around a six month interval)	Response process
Alternative-form reliability	Degree to which alternate forms of the same measurement instrument produce the same results	Response process
Questioning test takers	Interviewing respondents by	Response process

about the process of response to items	providing probing questions or allowing them to think-aloud as they respond to the items on an instrument to understand the process of response and its relationship with the intended construct	
Internal consistency (interitem) reliability	How well items reflecting the same construct yield similar results	Internal structure Consequences: absent in the traditional approach

Source: Adapted from Nunnally and Bernstein (1994), Ratanawongsa et al. (2008), and Trochim (2006)

Chapter Two

Review of the Literature

Method for Review of the Literature

The search strategy employed for this review of the literature involved three stages: (a) electronic search of literature databases, (b) hand search of leading medical education journals, (c) exploration in secondary statistical and research methods texts and statistical and research methods primary literature. These steps were designed to identify literature on reliability and validity in the field of medical education, reviews of validity evidence and/or factor analysis in medical education or related fields, and literature on best practices in establishing validity, including factor analysis methods.

First, ERIC, PsycINFO, and Medline databases were searched electronically with the dates 1999-2010. The year 1999 was selected as a cutoff since this was the year the *Standards of Educational and Psychological Testing* (AERA et al., 1999) was published which revised the framework for understanding validity evidence. Combinations of relevant keywords were applied within each database. Exact terms were identified using the thesaurus unique to each database which resulted in slightly different keywords for each database search. Specifically, the following terms were searched in ERIC – *validity, reliability, test construction, psychometrics, factor analysis, measures (individuals), medical schools, medical education, medical students,* and *review*. In PsycINFO, these terms were searched: *statistical validity, test validity, statistical reliability, test reliability, factor analysis, factor structure, measurement, psychometrics, medical education, medical students,* and *review.* In the CINAHL database, search terms included *reliability and validity, education (medical), factor analysis,* and *review*. In

Medline, search terms included *reproducibility of results, educational measurement, factor analysis (statistical), psychometrics, education (medical),* and *review*. The term *review* was used as a search term as well as publication type. Many articles were duplicated across searches and across databases due to crossover in search terms, and a large number of identified articles were instrument development articles appropriate for inclusion in the study, but not the literature review. Overall, few articles related to this literature review were identified. Thus, a hand search of leading medical education journals (e.g., *Academic Medicine, Medical Education, Advances in Health Sciences Education, Medical Teacher, Teaching and Learning in Medicine,* and *Journal of Continuing Education in the Health Professions*) seemed warranted. Keywords as described above were used to search electronically using each journal's search field, and titles and abstracts within recent issues were surveyed for relevance. In total, the various searches yielded well over 1000 articles; however, only 27 bear relevance to this research topic and meet the *Standards for Reporting on Empirical Social Science Research in AERA Publications* (AERA, 2006).

To inform the systematic review of instrument development in medical education, it was necessary to review both primary and secondary sources on factor analysis methods and techniques for establishing validity evidence. Sources were identified through a comprehensive review of reference lists of all systematic reviews included in the review of literature coupled with sources identified through previous work in this field. Each of these primary and secondary sources was reviewed to determine whether it might inform the development of the data extraction form specific to this study and subsequent appraisal of reported evidence.

This review of the literature offers first an overview of instrument development procedures. Second, techniques for establishing validity based on the framework of the *Standards for Educational and Psychological Testing* (AERA et al., 1999) are presented. As a popular method of establishing validity evidence, an overview of factor analysis methods is provided. This foundational information establishes best practices in this area; these best practices inform the review of literature and methodology for the current study. In addition, this information provides a foundation and context on which to frame the subsequent critique of previous systematic reviews of validity evidence and factor analysis literature presented at the end of this chapter.

Instrument Development

Within social sciences, psychometrics emerged as the field focused on theory and technique for measurement. At its inception, the focus was on ability testing which makes use of a classical measurement strategy known as item response theory (IRT). Over time, tenants of psychometrics were recognized as applicable to the measurement not only of ability but other psychological and social phenomena. Many of these phenomena involve constructs, also referred to as latent variables that cannot be directly observed or measured. Thus, additional measurement models developed to serve these efforts to measure and quantify latent variables using instruments such as assessments and questionnaires to measure uni- and multi-dimensional constructs.

Rigorous instrument development involves a series of six steps: (a) clearly define what is to be measured, (b) generate an item pool, (c) ask experts to review the item pool, (d) format and pilot test the items with a sample from the target population, (e) theoretically and empirically evaluate the items, and (f) revise items and establish optimal

scale length (AERA et al., 1999; DeVellis, 2003; Streiner & Norman, 2008). Proper implementation of each step should generate validity evidence to support inferences made based on the results from the instrument. Specifically, exploratory factor analysis is one leading, but methodologically complex, technique for establishing validity evidence through empirical evaluation of the fit of items to the construct being measured (Henson & Roberts, 2006; Kieffer, 1999). Detailed reporting of the instrument development process and accompanying validity evidence is an obligation of the developer; otherwise, thoughtful evaluation by the consumer is stifled.

History of Types of Validity. It is necessary to clarify the distinction between the contemporary understanding of validity evidence and the traditional classification system to frame the perspective adopted in this study.

Prior to the 1970s, efforts to validate instruments focused on the "three Cs", content validity, criterion validity, and construct validity (Cronbach & Meehl, 1955; Streiner & Norman, 2008). Each of these types of validity was seen as distinct from the other, and each required testing and validation to establish validity. From the traditional perspective, validity testing established an instrument as valid, which suggests an instrument might be valid or not valid. Two conceptual changes in the 1970s and 1980s upended the previous framework. First, a movement led by Cronbach (1971) emphasized that validity testing offered support not for the validity of the instrument but for the inferences made from an instrument in a given context with a given sample. Secondly, Messick (1975, 1980) asserted that the idea of types of validity was flawed. Rather, he purported that validity is a unitary concept for which supporting evidence helps establish the relationship between scores from an instrument and the construct. Therefore, validity

is seen as "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests" (AERA et al., 1999). These two notions were translated into recommendations for practice through the joint commission of the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education in the 1985 and the more recent 1999 *Standards for Educational and Psychological Testing*. Under the contemporary framework, validity is viewed as an argument made for the proposed interpretation of an instrument's scores based on an accumulation of evidence from five sources – test content, response process, internal structure, relationships with other variables, and consequences of testing (AERA et al., 1999). Which sources of evidence are most appropriate logically derive from the proposed interpretation and meaning of a given measure (Downing, 2003).

The terms reliability and validity are often paired in the measurement literature. Reliability does not imply validity; however, evidence of reliability is necessary for a strong validity argument. Like validity, reliability is not inherent to the instrument but reflects an interaction among the instrument, the specific participants, and the context in which the measurement occurs (AERA et al., 1999; Streiner & Norman, 2008). Generally, reliability is understood to refer to the consistency of scores on an instrument. This measure is essentially the ratio of "true" score variance to observed score variance. There are numerous types of reliability estimates, and their relevance and feasibility depend on the research design. For this study, sources of reliability evidence will be documented as they offer support for the five sources of validity evidence.

The *Standards* (AERA et al., 1999) contemporary framework is more than a decade old, but a full transition from the traditional classification of reliability and validity types has yet to occur as evidenced in the medical education literature (Artino, Durning, & Creel, 2010; Beckman et al., 2004; Hutchinson et al., 2002; Jha et al., 2007; Streiner & Norman, 2008; Tian et al., 2007; Veloski et al., 2005). Although efforts continue to be made to communicate validity as a unitary concept to medical education research practitioners (Andreatta & Gruppen, 2009; Artino et al., 2010; Cook & Beckman, 2006; Downing, 2004; Downing, 2003; Streiner & Norman, 2008), even some of these authors still preserve traditional validity terms (Artino et al., 2010; Streiner & Norman, 2008). As the preeminent source on this topic, the contemporary framework for validity evidence from the *Standards* (AERA et al., 1999) informs this study's research design enabling the comparison of current practices to best practices as defined by experts in this field.

Sources of Evidence for Validity.

Evidence Based on Test Content. From the beginning stages of instrument development, important validity evidence can be obtained. Evidence based on test content emerges in the development stages and reflects the relationship between items on the instrument and the construct of interest (AERA et al., 1999; Cook & Beckman, 2006; McMillan, 2008). To begin to evaluate content evidence, the construct to be measured must first be clearly defined (AERA et al., 1999; DeVellis, 2003; Streiner & Norman, 2008). This definition should reflect the theoretical underpinnings of the construct; however, in the absence of a strong theoretical basis, a tentative definition of the latent variable must be articulated to clarify what is being measured (DeVellis, 2003).

Once the boundaries of the latent variable are clearly delimited, a pool of items should be generated. The goal in item generation should be to cover all key concepts related to the construct, excluding items that are not directly related (DeVellis, 2003; Streiner & Norman, 2008). Multiple sources can be consulted to identify potential items including previous research, theory, expert opinion, direct observation, and interviews or focus groups with the target population (AERA et al., 1999; DeVellis, 2003; Streiner & Norman, 2008). If one is engaged in developing a new instrument, this suggests another instrument to measure the given construct was not available, not adequate, or not appropriate. However, items from existing instruments may be useful and offer the strength of already being tested. If new items need to be generated, the theoretical background used to define the latent variable should serve as a guide for key themes to include. It may be useful to observe individuals who engage in a particular behavior or present an attitude of interest to determine all elements of the construct. If observation is not practical, discussion with these individuals, through focus groups or key informant interviews, should generate key concepts of a given construct (AERA et al, 1999; DeVellis, 2003; Streiner & Norman, 2008). Finally, one should incorporate the use of expert opinion into any instrument development effort (AERA et al., 1999; DeVellis, 2003; Streiner & Norman, 2008). Experts in the construct under investigation can assist with item generation or review of the item pool to assess clarity, relevance, and thoroughness. Of particular importance is the evaluation of construct underrepresentation or irrelevance to ensure no critical areas are excluded or unrelated concepts included. (AERA et al., 1999; Downing & Haladyna, 2004). Efforts to develop a new instrument should include a combination of these sources. Subsequent reporting, including detailed

description of the experts and the procedures used to define the construct and develop and refine items, would highlight this evidence based on test content.

Evidence Based on Response Process. Analysis of the response process of participants engaged in a pilot study or formal administration of an instrument can provide further validity evidence by supporting the fit between the construct of interest and the response process engaged in by the participants (AERA et al., 1999; Downing, 2003). Observations of participants in performance based outcome measures, records documenting phases of the development of a written response, or results from questioning participants about their response to particular items either during or after administration of the instrument are valuable ways to understand the response process and its relationship to the construct (AERA et al., 1999; Cook & Beckman, 2006; Downing, 2003). In addition to analyzing the response process of the participant, evaluation of the process engaged in by raters or scorers – how well they apply particular criteria in rating or scoring – is also important where relevant (AERA et al., 1999).

Following administration of an instrument in development, several reliability measures are available for analysis of the consistency of scores in light of a single source of error within the measurement and response process. At the participant level, empirical analysis of consistency across time (e.g., test-retest reliability and test-retest stability) is available for application. Inter-rater reliability and intra-rater reliability provide evidence for the consistency of scoring across multiple raters or for the same rater across multiple occasions, respectively. Each of these methods are quite popular; however, generalizability theory (GT) (Cronbach, Glesler, Nanda, & Rafaratnam, 1972) is more powerful and allows the researcher to parse out variance for all sources of error at once

and determine each source's influence on the measurement process (AERA et al., 1999; DeVellis, 2003; Downing, 2004; Streiner & Norman, 2008). However, GT is based on a random ANOVA model with strong methodological assumptions that are often unmet in social, behavioral, and educational studies; therefore, GT is not widely adopted in psychometric studies (Streiner & Norman, 2008).

Evidence Based on Internal Structure. Empirical analysis in light of the conceptual framework for the construct of interest is critical for evaluation of the instrument and offers evidence for internal structure. Internal structure, as a source of validity evidence, refers to the degree to which the relationships between items or between underlying factors are consistent with the construct of interest (AERA et al., 1999). As Downing (2003) describes it, "scores on test items or sets of items intended to measure the same variable, construct or content area should be more highly correlated than scores on items intended to measure a different variable, construct, or content area" (p. 834). Generally, both internal consistency reliability and factor analysis data are considered sources of internal structure evidence as the first speaks to the homogeneity of test items and the second to the internal structure of the test. Further, consistency across equivalent measures (e.g., alternative or parallel forms reliability) may be thought of as weak evidence, relative to factor analysis, for internal structure.

Internal consistency, a measure of reliability, enables a very accessible empirical investigation of the correlations between items and sets of items based on a single administration of the instrument. Kuder-Richardson 20 (1937) and Cronbach's (1951) coefficient alpha are two methods that provide an average of all possible split-half reliabilities for an instrument. KR-20 applies to dichotomous item responses whereas,

Cronbach's alpha is used for items with more than two response options. Although Cronbach's alpha is often applied, McDonald (1999) offers two justifications for his recommendation for calculating coefficient omega in lieu of alpha for factor analysis studies that suggest a multidimensional instrument. First, alpha tends to underestimate reliability compared to omega. Second, summation of a total score for multidimensional instruments is inappropriate, limiting the use of alpha to measurement of internal consistency at the factor level. However, in these circumstances, omega may still be applied to calculate an overall reliability coefficient. Overall, measures of internal consistency should be interpreted with caution as they fail to account for multiple potential error sources such as time and different raters and should be combined with other reliability measures (AERA et al., 1999; Streiner & Norman, 2008).

Factor analysis provides the capacity to explore and test for evidence of the dimensionality of a construct (Cook & Beckman, 2006; DeVellis, 2003; Streiner & Norman, 2008). Thus, whether a construct is defined as uni- or multi- dimensional, factor analysis can provide statistical evidence of how well patterns of responses conform to the construct as defined. Because factor analysis is one of the most useful, but complex, techniques for establishing validity evidence based on internal structure, the methodological steps involved will be reviewed in more detail in the following section.

For assessment instruments, differential item functioning (DIF) may serve as an additional technique to explore evidence for validity. According to the *Standards for Educational and Psychological Testing*, differential item functioning "occurs when different groups of examinees with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular

item" (AERA et al., 1999, p.13). It should be noted that in some instances evidence of DIF may not be detrimental to the argument for validity if, based on the conceptual framework, the variations in performance can be explained due to specific test content or task (AERA et al., 1999).

Evidence Based on Relationships with Other Variables. Additional data collection from participants on other instruments or outcome measures presents opportunities to investigate validity based on relationship with other variables. Informed by the construct, these other variables may be expected to be related or unrelated to scores on the instrument in development, or it may be hypothesized that scores are predictive of some other variable(s). Under the traditional framework and still in some current writings, predictive, concurrent, convergent, discriminant, and divergent validity are referenced as types of validity related to this contemporary source of validity evidence (DeVellis, 2003; Streiner & Norman, 2008). Instead, from a contemporary perspective, terminology like convergent and discriminant evidence and test-criterion studies is employed (AERA et al., 2009; Downing, 2004; McMillan, 2008). Convergent evidence shows a positive correlation between scores on the instrument and scores on another instrument or outcome measure intended to measure the same construct. On the other hand, discriminant evidence would show a low or no correlation between scores on the instrument and a conceptually different outcome measure. The multitrait-multimethod matrix is a classic design used to demonstrate these two types of evidence based on relationships with other variables (Campbell & Fiske, 1959). Test-criterion evidence relates to an essential question "How accurately do test scores predict criterion performance?" (AERA et al., 1999, p. 14). These studies may be referenced as predictive

or concurrent studies which are differentiated by timing of the measures. Predictive criterion reference to the future; concurrent criterion are measured simultaneously with the instrument in development. It should be noted test-criterion relationships are only as strong as the reliability and validity of the inferences from the criterion measure (AERA et al., 1999). Divergent validity suggests the "ability of a measure to yield different mean values between relevant groups" (Nunnally & Bernstein, 1994, p. 6).

Evidence Based on Consequences of Testing. A new way of conceptualizing validity, evidence based on test consequences is not well addressed in the previous validity framework. The evidence for this source of validity has been considered more subjective than others (Downing, 2003), and thus is still a controversial topic in validation (Cook & Beckman, 2006). Although many instruments are used for solely research purposes or formative feedback and remediation, for those used to make high stakes decisions, it is imperative to ensure "the desired results were achieved and unintended effects avoided" (Cook & Beckman, 2006, p.166.e12; Downing, 2003; Messick, 1975; Messick, 1980; Streiner & Norman, 2008). To support this type of validity evidence, researchers should describe clearly the process of scoring, report cut-off scores applied and justify these scores, calculate and report classification accuracy when relevant, and report the standard error measurement (AERA et al., 1999; Downing, 2003). In addition, instrument developers should look to outcomes caused by the assessment. Positive and negative, as well as intended and unintended, consequences of testing should be reviewed to ensure fairness and minimize bias (Andreatta & Gruppen, 2009; AERA et al., 1999). For example, if a screening tool for high cholesterol helps physicians place patients into treatment groups that lead to lowered cholesterol, then this

would be supportive evidence. On the other hand, if this screening tool was found to differentiate patients on a characteristic unrelated to the construct of high cholesterol such as race or gender, then this would be reason for concern about the validity of placement into treatment based on the screening tool. From an educational assessment perspective, this would be termed differential test functioning (DTF), or the evaluation of whether sets of items function differently for different groups (Badia, Prieto, & Linacre, 2002).

Combining evidence based on each of these five sources can lead to a well-developed argument for the reliability and validity of inferences made from an instrument designed to measure a certain construct. However, thoughtful planning by the researcher is required to ensure rigorous instrument development methods are employed and thus supportive validity evidence available. It is the test consumer's responsibility to evaluate whether interpretations to be made from an instrument are sufficiently trustworthy. However, it is incumbent on the test developer to clearly describe the methods and report the evidence based on test content, response process, internal structure, relationships to other variables, and test consequences to enable such an evaluation.

Factor Analysis. Factor analysis is a useful technique for establishing validity evidence based on internal structure in instrument development. As mentioned earlier, it provides empirical evidence of the dimensionality of a construct. Factor analysis is useful and often applied in medical education instrument development; however, it is methodologically complex in comparison to other techniques for establishing validity making it vulnerable for misuse (Henson & Roberts, 2006). The procedure involves a series of methodological steps, each requiring informed decision making by the researcher, as different approaches can yield distinctly different results that can impact

inferences made from an instrument (Henson & Roberts, 2006; Kieffer, 1999). Given the number of techniques available in factor analysis design, it is critical for the researcher to clearly report each step and provide support for why specific choices were made. This enables evaluation of the research design and the potential for replicability. From the literature derive five necessary elements of factor analysis that should each be thoughtfully planned, reported, and justified: (a) model of analysis, (b) sample size criteria, (c) method of extraction, (d) rotation method and (e) criteria for factor retention (Comrey & Lee, 1992; Floyd & Widaman, 1995; Gorsuch, 1983; Reise et al., 2000; Tabachnick & Fidell, 2007). Each element will be discussed separately.

Model of Analysis. Exploratory factor analysis (EFA) and principal components analysis (PCA) are often used interchangeably; however, the two are distinctly different models of analysis (Bentler & Kano, 1990; Floyd & Widamen, 1995; Gorsuch, 1990; Mulaik, 1990; Reise et al., 2000; Snook & Gorsuch, 1989; Widamen, 1990, 1993, 2007; Tabachnick & Fidell, 2007). EFA, also referred to as the common factor model, seeks to identify the latent variables, referred to as factors, which explain the correlations between the observed variables. The hypothetical latent variable is understood to determine the scores on the observed variables. For PCA, the components identified through the analysis are not latent variables but represent linear combinations of the observed variables; the components are weighted sums of item responses. The key difference between the two models lies in the mathematical equation underlying each technique. EFA aims to explain only shared or common variance; whereas, PCA attempts to explain the total variance (Costello & Osborne, 2005; DeVellis, 2003; Floyd & Widamen, 1995; Tabachnick & Fidell, 2007; Widamen, 1993). Thus, the correlation or covariance matrix

on which the analysis is performed differs between the two models. For PCA, the goal is essentially data reduction, and all variance – common, unique, and error – is maintained in the correlation or covariance matrix on which the analysis is based. EFA seeks to estimate an error-free factor solution, thus analysis is limited to common variance shared between observed variables. Variance unique only to an individual variable and error variance are parceled out of the equation. Therefore, EFA is based on a correlation or covariance matrix that includes only common variance.

Empirical research using both real and simulated data sets has produced instances when EFA and PCA lead to similar results (Velicer & Fava, 1998; Velicer & Jackson, 1990a; Velicer, Peacock, & Jackson, 1982). A number of researchers support these findings and purport differences between EFA and PCA are minimal and have little practical impact on the interpretation of results (Guadagnoli & Velicer, 1988; Schoenmann, 1990; Steiger, 1990; Velicer & Jackson, 1990b; Zwick & Velicer, 1986). However, the data sets applied in these studies were limited to strong, quality data with high saturation (i.e., large observed variable to factor ratios) and strong factor loadings. Follow-up studies applying varied quality of data along the previously listed dimensions found important differences in results between EFA and PCA (Snook & Gorsuch, 1989; Widamen, 1990). Specifically, PCA overestimated factor loadings with overestimation worsening for higher communalities and fewer variables per factor (Snook & Gorsuch, 1989; Widamen, 1990); whereas, EFA did not produce bias in factor loadings across samples with different data quality (Widamen, 1990). In addition, PCA remains directly linked to the original data set, including its error variance term, limiting the potential for replication (Mulaik, 1990). On the other hand, EFA estimates an error-free model that

should enable replication studies and hypothesis testing based on underlying variables and should generalize better than PCA to confirmatory factor analysis models (Floyd & Widamen, 2005; Mulaik, 1990). With the right design, differences between the two procedures may be minimized; however, in this ongoing debate, there is much support for the limited use of PCA for data reduction or summarization and endorsement of EFA for instrument development (Bentler & Kano, 1990; Costello & Osborne, 2005; Floyd & Widamen, 1995; Gorsuch, 1990; Mulaik, 1990; Snook & Gorsuch, 1989; Widamen, 1990, 1993, 1997). An understanding of these differences between the two models highlights the importance of reporting the model of analysis in factor analysis research literature to inform the reader. Also, this illustrates the need for thoughtful, informed researchers who are able to select the appropriate model based on the research question.

Sample Size Criteria. There is a lack of consensus on ideal sample size for factor analysis research; though, in general, factor solutions from larger samples tend to produce more precise estimates of the population and to be more stable across sampling (DeVellis, 2003; MacCallum, Widamen, Zhang, & Hong, 1999). Rules of thumb are plentiful and reference both overall sample sizes as well as participant to variable ratios. Recommended participant to variable ratios include a range – 3-6:1 (Cattell, 1978), 5:1 (Gorsuch, 1983), 5-10:1 (Tinsley & Tinsley, 1987), 10:1 (Everitt, 1975; Costello & Osborne, 2005). Other researchers purport a minimum overall sample size, like Tabachnick and Fidell (2007) who recommend at least 300 participants. A popular metric for evaluating sample size was proposed by Comrey and Lee (1992) and indicates a sample of 100 is poor, 200 is fair, 300 is good, 500 is very good, and 1000 or more is excellent. Evidence suggests no recommendation for total sample size or participant to

variable ratio will be appropriate for all factor analysis studies (Guadagnoli & Velicer, 1988; Hogarty, Hines, Kromrey, Ferron, & Mumford, 2005; MacCallum & Tucker, 1991; MacCallum et al., 1999). Specifically, MacCallum, Widamen, Zhang, and Hong (1999) commented that "common rules of thumb regarding sample size in factor analysis are not valid or useful" (p. 96). Together, these studies clarify that factor solutions may be negatively influenced by a small sample size when data quality is low (e.g., low communalities, low saturation); however, the quality of factor solutions improves as communalities and saturation improve, making overall sample size less important. Although some researchers may know what communalities or the number of variables per factor to expect prior to performing the factor analysis, most researchers will not. Thus, MacCallum and colleagues (1999) suggest using as large a sample as possible and then applying these quality criteria after the factor analysis to evaluate sample size and its influence on the factor solution.

Method of Extraction. The distinction between PCA and EFA refers to the model of analysis; however, within the EFA common factor model, there are several methods of extraction of which maximum likelihood, principal axis factoring, and generalized least squares seem to be most frequently employed. Maximum likelihood (ML) makes use of a statistical criterion to determine the number of factors to extract (DeVellis, 2003). ML applies the χ^2 goodness-of-fit statistic to test the null hypothesis of no discrepancy between the observed and predicted correlation or covariance matrices. This method assumes multivariate normality; therefore, this assumption should be tested prior to analysis (Costello & Osborne, 2005; Floyd & Widamen, 1995; Tabachnick & Fidell, 2007). As with other tests of significance, the ML χ^2 goodness-of-fit test is

sensitive to sample size. As sample size increases, the researcher should be cautious about potential overestimation of the number of factors (Floyd & Widamen, 1995). Principal axis factoring (PAF) is commonly supported for data that are not normally distributed (Costello & Osborne, 2005; Floyd & Widamen, 1995; Tabachnick & Fidell, 2007). Generalized least squares (GLS) offers an extraction method suitable for categorical data (Norris & Lecavalier, 2010). The distinction made for data type is an important one; the level of measurement for the observed variables (i.e., items) should be the primary criterion used to select an extraction method. For an instrument with all continuous variables, EFA-ML is recommended (Muthen & Muthen, 2010). Weighted least squares factor analysis (EFA-WLS), a special case of GLS, should be used for ordinal level items (Dumenci & Achenbach, 2008; Muthen & Muthen, 2010). Dumenci and Achenbach (2008) studied the effects of estimation method on factor scores from ordinal data from uni-dimensional Likert scale instruments. They found both PCA and EFA-ML extraction methods led to biased factor scores. The bias was noted particularly at the ends of the total score range. They suggest this issue can be resolved through application of EFA-WLS that accounts for the ordered nature of Likert scale items. Lastly, EFA-MLR should be used for instruments with non-normal, continuous item distributions (Muthen & Muthen, 2010).

Method of Rotation. Use of rotation in factor analysis will often enhance interpretability of the factor structure by seeking to maximize simple structure; simple structure implies each variable has only one high factor loading and all other low or zero loadings (Browne, 2001; Thurstone, 1947). There are two major categories of rotation from which a researcher might select a specific rotation method: orthogonal and oblique.

Orthogonal rotations do not allow factors to correlate; whereas, oblique rotations do allow correlation between factors (DeVellis, 2003; Reise et al., 2000). Quartimax and varimax are the main orthogonal rotations. Quartimax is less popular because of its tendency to produce a general factor (Comrey & Lee, 1992; Gorsuch, 1983; McDonald, 1985), "one factor with all major loadings and no other major loadings in the rest of the matrix, or have the moderate loadings all retained on the same factor" (Gorsuch, 1983, p. 184). Varimax rotation is the most popular rotation procedure currently and is the default method in most statistical software programs (Comrey & Lee, 1992; Henson & Roberts, 1996; Widamen, 2007). Direct oblimin and promax are generally recognized oblique rotations, with promax better supported (Gorsuch, 1983; McDonald, 1985).

DeVellis (2003) suggests researchers use existing theory to inform selection of an appropriate rotation method based on if and to what extent factors are correlated. Other methodologists suggest oblique rotations fit better conceptually with most social science constructs under measurement (Norris & Lecavalier, 2010; Raykov & Marcoulides, 2011; Reise et al., 2000; Tabachnick & Fidell, 2007) and provide additional information on the relationship between factors that may enhance understanding of the construct (Norris & Lecavalier, 2010). In addition, if an oblique rotation suggests factors are not correlated, then the orthogonal rotation may instead be interpreted (Raykov & Marcoulides, 2011; Reise et al., 2000; Tabachnick & Fidell, 2007). Although oblique rotations may offer conceptual advantages, orthogonal rotations remain the default in most statistical packages (Henson & Roberts, 2006; Widamen, 2007), and researchers often employ orthogonal rotations based on a perceived ease of interpretability (Reise et al., 2000). Regardless of which rotation method is applied, Floyd and Widamen (1995) emphasize

the importance of complete reporting in factor analysis studies, including the rotation method, justification for the rotation method, and appropriate matrices, as described below.

For orthogonal rotations, only a factor loading matrix must be interpreted and reported; each factor loading represents the "extent of the relationship between each observed variable and each factor…the loading matrix [is interpreted] by looking at which observed variables correlate with each factor" (Tabachnick & Fidell, 2007, p. 609). However, oblique rotations include more complexity with a factor correlation, structure, and pattern matrix. The factor matrix indicates correlations between factors. The structure matrix presents correlations between factors and observed variables. Finally, the pattern matrix, which is used for interpretation, presents the unique relationships between factors and observed variables. Both the factor correlations and pattern matrix should be reported in factor analysis instrument development literature (Floyd & Widamen, 1995; Tabachnick & Fidell, 2007).

Criteria for Factor Retention. Once factors have been extracted, researchers must decide how many factors to retain in the factor solution. A number of decision rules and criteria are available to address this methodological decision step in factor analysis, each with more or less potential for accuracy.

One of the first decision rules was proposed by Kaiser (1960) and purports factors with eigenvalues greater than one should be retained. An eigenvalue represents the amount of variance captured by the individual factor; values greater than one indicate the factor explains more variance than one single item. On the other hand, factors with values less than one fail to explain even as much variance as one item adding little value

to the model (DeVellis, 2003). The eigenvalue greater than one rule is quite popular in practice and is currently the default criterion in most statistical software packages (Comrey & Lee, 1992; Widamen, 2007; Zwick & Velicer, 1986); however, many argue this decision rule is the least accurate often leading to extraction of too many factors (DeVellis, 2003; Floyd & Widamen, 1995; Reise et al., 2000; Zwick & Velicer, 1986). Specifically, Zwick and Velicer (1986) found, in their comparison of five criteria for factor retention, the eigenvalue rule overestimated the number of factors with overestimation worsening as the number of variables increased. Exclusive reliance on this criterion is not recommended.

The scree test, articulated by Cattell (1966), represents a second popular criterion for determining the number of factors to retain. The scree test plots the eigenvalues of each factor in descending order on a chart where the factors are placed on the x-axis and the eigenvalues on the y-axis. The factors on the vertical slope are retained as valuable factors, and those factors on the horizontal are considered the scree (or rubble at the bottom of the mountain) and discarded (Comrey & Lee, 1992; DeVellis, 2003). For PCA, this bend in the slope, or elbow, will often occur at the eigenvalue equal to 1.0 mark; however, for EFA, there may be an unclear or multiple bends (Comrey & Lee, 1992; Floyd & Widamen, 1995; Gorsuch, 1983; Zwick & Velicer, 1986). This method can be perceived as subjective, though Zwick and Velicer (1986) found it to be less variable than the eigenvalue rule and inter-rater reliability between two raters was moderate. When there were inaccuracies, like the eigenvalue rule, the scree test tended to overestimate the number of factors (Zwick & Velicer, 1986).

Parallel analysis, a third potential criterion for factor retention, is essentially a sophisticated extension of the scree test (Horn, 1965). Using the same number of participants and variables as the real data set, random data sets are generated. The scree plot of eigenvalues for the random data set is plotted against those of the real data set. The point where the two curves cross is established as the cut-off point; thus, no real data factors are retained that explain less variance than factors from the random data. Zwick and Velicer (1986) found parallel analysis to be the most accurate and least variable criterion; however, most researchers do not have access to this calculation through common statistical software packages (Norris & Lecavalier, 2010).

Less common statistical criteria include the Bartlett's test and minimum average partial. Bartlett's test is similar to the scree test, evaluating the quality of the remaining factors; however, it is sensitive to sample size, the number of variables, and factor saturation (Zwick & Velicer, 1986). Minimum average partial was found to be more accurate than the eigenvalue rule, scree test, and Bartlett's test. In minimum average partial, as each factor is extracted from the matrix, a partial correlation matrix that includes the remaining variance is calculated. Essentially, factors continue to be extracted until all common variance is represented in the extracted factors and only unique variance remains in the matrix (Bandalos & Boehm-Kaufman, 2009). Unlike other methods, it tends to underestimate the number of factors to extract by ignoring minor components (Zwick & Velicer, 1986). From these available statistical criteria, use of parallel analysis and the scree test in conjunction is recommended (Zwick & Velicer, 1986).

Other criteria may also be applied in conjunction with the above-mentioned statistical approaches to determine the number of factors to retain. Some researchers may use the percent of explained variance in a factor solution to support the number of retained factors. Floyd and Widamen (1995) suggest as a minimum standard that 80 percent of common variance be explained by the factor solution; however, a commonly accepted minimum was not identified. Although it is unclear what minimum should be employed, the percent of explained variance for each factor prior to rotation and the percent of explained variance for the whole solution after rotation should always be reported to inform the reader (Floyd & Widamen, 1995).

Factor saturation, or the number of high loading items on a factor, can also be used to determine whether a factor should be retained. Support can be found for a minimum of three items per factor; less than three may suggest an unstable factor (Floyd & Widamen, 1995; Tabachnick & Fidell, 2007). Recommendations for a minimum factor loading for an item to load on a given factor vary (Norris & Lecavalier, 2010). Comrey and Lee (1992) suggest a scale of quality of factor loadings that is often referenced: .71 is excellent, .63 is very good, .55 is good, .45 is fair, and .32 is poor. On the other hand, Tabachnick and Fidell (2007) more recently purport a minimum of .32 is acceptable. Overall, the choice of factor loading is at the researcher's discretion, and if homogeneity of responses is expected in the data, lower loadings should be interpreted (Comrey & Lee, 1992). Overall, researchers are encouraged to use multiple criteria translated in view of prior theory and interpretability (Floyd & Widamen, 1995; Norris & Lecavalier, 2010).

As evidenced here, factor analysis is a very useful, but complex technique for establishing evidence for validity based on internal structure with numerous steps and methodological decision points. This illuminates the importance of clear and complete reporting by researchers in order for the reader to understand the details and quality of the factor analysis performed.

Reviews of Validity Evidence

In medical education research, reviews have examined the reliability and validity evidence reported in studies, but this is not typically the exclusive focus of the review. Rather the evaluation of psychometric reporting practices is often paired with a primary research question related to the availability of certain outcome measures, research designs in which these measures are applied, and/or quality of the research process more broadly. However, some of the findings are still relevant to our understanding of the techniques for establishing validity applied in medical education research. Relevant findings are reviewed here.

Organization of previous reviews in medical education reflects several approaches to understanding validity. In Tian and colleagues' (2007) review of continuing medical education (CME) evaluation studies (n = 32), the validity framework applied was not explicated. Though a tertiary finding for their study, results indicate of the ten studies that developed and applied a new instrument, none reported reliability or validity evidence.

Using standards supported by the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPICC), Jha and colleagues (2007) reviewed measures of medical student attitudes toward professionalism (n = 97). They found approximately

half of the studies reported both reliability and validity evidence, though specific techniques for validation were not elaborated. Although 53 percent of these studies reported the theoretical framework informing the test content, very limited information was provided on item development and review.

Several reviews applied the traditional validity framework to extract specific types of reliability and validity evidence from the literature (Beckman et al., 2004; Hutchinson et al., 2002; Veloski et al., 2005). Based on a review of both instruments in development or testing stages for assessment in postgraduate medical certification (n = 55), Hutchinson and colleagues (2002) found inter-rater reliability and internal consistency reliability were most often reported, with little evidence for construct validity. Beckman's research team (2004) conducted a review of instruments for evaluating clinical teaching (n = 21). They found internal consistency to be the most employed psychometric measure and found the consistent use of expert review of test content. Veloski and colleagues (2005) reviewed articles reporting on measures of student and resident professionalism (n = 134). Although the traditional framework was used to extract data from the studies, coders were asked to evaluate whether the reliability and/or validity evidence met the *Standards for Educational and Psychological Testing* (AERA et al., 1999). Their findings are consistent with other reviews with internal consistency, inter-rater, and test-retest reliability most often reported; however, roughly half of the articles failed to report any reliability evidence. One-third provided no validity evidence, and of the others, most reported expert review for content validity. Using a five point Likert scale, coders rated the quality of the reliability and validity

evidence in light of the *Standards* (AERA et al., 1999); only 15 of 134 were rated as high or very high.

Three reviews were identified through the review of literature that applied the contemporary framework for validity evidence as espoused in the *Standards* (AERA et al., 1999). First, in a review of instruments used for evaluation of evidence-based practice (n = 115), Shaneyfelt and colleagues (2006) found the majority of studies reported at least one source of validity evidence, but only 10 percent used multiple types of validity evidence to support inferences made from the instrument results. Unlike the previous findings, most validity evidence was based on relationships to other variables, followed by evidence based on test content and internal structure.

Second, Lubarsky and colleagues (2011) conducted a review of articles related to script concordance testing (SCT) to evaluate the validity evidence available to support this specific assessment method. The number of reviewed articles is unclear; however, the authors indicate evidence based on test content and internal structure measured using internal consistency reliability as most prevalent. Only a few articles reported on evidence based on relationships with other variables, and evidence based on response process and consequences of testing was particularly weak.

Finally, Ratanawongsa and colleagues (2008) reviewed evaluations of CME, limited to randomized control trial (RCT) and historic/concurrent comparison designs. It is important to note, they only included studies that reported either reliability or validity evidence, narrowing their review from 136 studies to 47 studies. They then made their unit of analysis the instrument, rather than the overall study, as more than one instrument was included in some studies. Thus, of 62 reviewed instruments, only 16 percent

reported both reliability and validity evidence. Validity evidence was reported in half of the studies and mostly involved a description of experts engaged in the review of test content. The majority of studies reported some evidence based on internal structure measured by internal consistency or based on response process and measured by inter-rater reliability. None of the authors included evidence based on test consequences.

To execute their review, Ratanawongsa et al. (2008) extracted data based on the traditional framework for validity and fitted these data to the contemporary framework (e.g., test-criterion validity coded as evidence based on relations to other variables, internal consistency coded as internal structure evidence). They felt they needed to extract the data based on how it would be presented in the articles and acknowledged that most in medical education do not have a full understanding or have not yet adopted the contemporary framework with validity as a unitary construct. This approach will inform the data extraction process of the current study.

The consensus across these findings suggests researchers provide limited evidence for reliability and validity of measures, constraining the instrument consumer's capacity to make informed selection of measures for use in their own educational practice and research. Although each of these reviews provides valuable information to enhance the understanding of reporting of reliability and validity evidence in medical education, there are limitations. Each review is narrowly focused on a subset of the medical education research literature delimited by a point on the continuum or measures of a particular construct. Most are not exclusive to instrument development. In addition, few reviews have applied the *Standards* (AERA et al., 1999) as an organizational framework for evaluating reported evidence. Thus, a comprehensive review of the application of

techniques for establishing validity in instrument development articles, informed by the *Standards* (AERA et al., 1999) contemporary framework for validity evidence, is necessary.

Reviews of Factor Analysis

Reviews of factor analysis procedures can be found in the psychology literature, each systematically appraising either a particular specialty area of psychology or particular research journals. Here, the scope of each review from psychology is presented followed by a synthesis of findings across the reviews. Reviews of factor analysis in education are fewer and are discussed after those from psychology.

Reviews of Factor Analysis in Psychology. The most notable and frequently cited review by Fabrigar and colleagues (1999) evaluated factor analysis articles (*n* = 159) published 1991-1995 in the *Journal of Personality and Social Psychology* and the *Journal of Applied Psychology.* Park et al. (2002) replicated the design of Fabrigar et al. (1999) and conducted a review of communication research factor analysis articles (*n* = 119) published 1990 to 2000. They limited their search to three communication journals, *Human Communication Research, Communication Monographs,* and *Communication Research* (N = 119). Norris and Lecavalier (2010) narrowed their review to focus on the developmental disabilities field within psychology. Specifically, they reviewed factor analysis articles (*n* = 66) from five developmental disability journals – *American Journal on Mental Retardation, Journal of Autism and Developmental Disorders, Journal of Intellectual Disability research, Journal of Intellectual and Developmental Disabilities,* and *Research in Developmental Disabilities* – published January 1997-May 2008. Finally, Worthington and Whittaker (2006) expanded their review of factor analysis

including articles employing both exploratory factor analysis and confirmatory factor analysis. They focused specifically on studies (n = 23) published in the *Journal of Counseling Psychology* from 1995-2004.

The systematic review design across these studies focused on evaluation of four key factor analysis methodological decisions – model of analysis, sample size, rotation method, and criteria for factor retention. Findings suggest at least half of factor analysis studies applied PCA, roughly one-third failed to articulate the model of analysis, and the remainder used EFA (Fabrigar et al., 1999; Norris & Lecavalier, 2010; Park et al., 2002; Worthington & Whittaker, 2006). PCA was often inappropriately applied when the research questions were not focused on data reduction but on exploring underlying dimensions of a construct. Worthington and Whittaker (2006) note the use of PCA in the earlier studies reviewed and EFA in the latter studies and suggest perhaps a trend away from PCA, though that finding is not confirmed in Norris and LeCavalier's (2010) later work. Evidence was found for the widespread use of adequate to large sample sizes in the factor analysis study designs (Fabrigar, et al., 1999; Norris & Lecavalier, 2010; Park et al., 2002). Orthogonal varimax rotation was the most often selected rotation method (Fabrigar et al., 1999; Norris & Lecavalier, 2010; Park et al., 2002; Worthington & Whittaker, 2006), in spite of instances with clear theoretical or empirical evidence to suggest high correlations between factors warranting an oblique rotation (Norris & Lecavalier, 2010; Worthington & Whittaker, 2006). Approximately 20 percent of authors did not report the rotation method (Fabrigar et al., 1999; Park et al., 2002), and few provided justification for the selected method (Worthington & Whittaker, 2006). Most reviews found factor analysis researchers made use of multiple criteria for determining

the number of factors to retain, with the eigenvalue greater than one rule, scree test, and meaningfulness or interpretability most often applied (Fabrigar et al., 1999; Norris & Lecavalier, 2010; Park et al., 2002; Worthington & Whittaker).

Norris & LeCavalier (2010) expanded on the methodological decisions previously reviewed and thus offer additional information to the understanding of factor analysis in psychology. Specifically, their findings indicate roughly 40 percent of studies did not report the required minimum value for an item to load on a factor. In addition, although half of the studies reported the full factor loading matrix, approximately one-quarter did not present any factor loadings, and the remaining one-quarter only reported loadings that met or exceeded the required factor loading magnitude.

The consensus across these systematic reviews suggests some inappropriate use of factor analytic methods, particularly PCA over EFA and orthogonal over oblique rotations. In addition, their results indicate the frequent failure to report methodological decisions required for other researchers to evaluate and potentially replicate the analysis. Though related within the social sciences, these reviews are limited to psychology and may not reflect work in education.

Review of Factor Analysis in Psychology and Education. Henson and Roberts (2006) offer a review not exclusive to either psychology or education. Fifteen applications of exploratory factor analysis were selected from each of four journals, *Educational and Psychological Measurement*, *Journal of Educational Psychology*, *Personality and Individual Differences*, and *Psychological Assessment*, resulting in a review of 49 articles published prior to the year 2000. Again, sample sizes were generally acceptable. Slightly more than half of the factor analysis studies used PCA,

roughly 20 percent used EFA, and nearly 15 percent did not report their model of analysis. Reflecting other findings, orthogonal varimax rotation and the eigenvalue greater than one rule and scree test factor retention criteria were most often applied. Henson and Roberts' (2006) findings differ from other reviews in finding that the majority (55%) of studies applied only one criterion in determining the number of factors to retain. Omission in reporting of critical methodological decisions in these studies creates questions about research quality. Although Henson and Roberts (2006) did not review CFA studies, they did assess whether a CFA was warranted in place of EFA and found one-third of studies failed to implement a CFA when appropriate and provided no justification for this design decision.

Reviews of Factor Analysis in Education. Finally, Pohlmann (2004) and Henson, Capraro, and Capraro (2004) conducted the only identified reviews of factor analysis exclusive to education. Pohlmann (2004) reviewed principal component analysis and exploratory and confirmatory factor analysis studies (n = 25) published 1992-2002 in *The Journal of Educational Research*. Of the 25 studies, nine employed PCA, nine EFA, three CFA, and four did not identify the model. Again, varimax was the most common rotation. Different from previous reviews, prior theory as a guide for factor retention was cited most often, followed by the eigenvalue greater than one rule and scree test. The second review of educational factor analysis studies by Henson and colleagues (2004) included review of 49 EFA and PCAs from three education journals – *American Educational Research Journal, Journal of Educational Research,* and *The Elementary School Journal*. As previously found, sample sizes tended to be large. One-third of the studies applied PCA, one-third did not identify the model of analysis, and the remainder

used EFA. This is consistent with previous findings where PCA is used as often as or more often than EFA (Fabrigar et al., 2006; Henson & Roberts, 2006; Norris & Lecavalier, 2010; Park et al., 2002; Worthington & Whittaker, 2006). Applying an *a priori* statement of the number of factors to retain was given most often as the criterion for retention of factors. Otherwise, the eigenvalue greater than one rule and scree test were most often used. Overall, the majority of studies applied only one factor retention decision rule, and one-quarter did not report this information. Unlike other studies, oblique and orthogonal rotations were almost equally employed (40.8% and 34.7%, respectively). Similar to Henson and Roberts' (2006) review of educational and psychological factor analysis, in this review Henson and colleagues (2004) also explored whether CFA was appropriate in any of the research designs and found one-third of studies failed to employ CFA when warranted. They also investigated additional methodological decisions finding most studies failed to report the eigenvalues for factors retained, and more than half did not report the variance explained by the factor solution.

Findings from these three reviews that include educational factor analysis studies are generally consistent with results from reviews within psychology. Overall, evidence suggests researchers do not consistently meet best practices in conducting factor analysis and reporting on methodological decisions. Though including some educational research, these reviews did not include medical education factor analysis research studies. Schonrock-Adema and colleagues (2009) articulate within the medical education research community the need for improvement in the use of factor analysis. Although their recommendations are based on best practices from the literature, they are not informed by current factor analysis practice by medical education research practitioners. To date,

reviews of the literature have not identified a review of factor analysis in medical education. Given concerns about factor analysis research practice in related fields, such a review appears warranted.

This review of the literature offers an overview on establishing validity evidence through rigorous instrument development employing factor analysis and demonstrates the complexity and diversity of options within these procedures. Though best practices have been articulated, effective implementation requires an informed, thoughtful researcher who can apply and report best practices in instrument development research. Limited evidence from medical education and supplemental evidence from psychology and education more generally suggest gaps in translating factor analysis best practices and the *Standards for Educational and Psychological Testing* (AERA et al., 1999) into research. However, a comprehensive review of the extent to which instrument development in medical education complies with these best practices remains relatively unclear. This study aims to address this gap.

Chapter Three

Methodology

The purpose of this study was to conduct a comprehensive review of instrument development articles employing exploratory factor analysis or principal component analysis published in medical education from 2006 through 2010. This review enabled the description and assessment of the reporting of methods and validity evidence. Findings from this study inform the following two research questions.

Within medical education instrument development literature, including undergraduate, graduate, and continuing medical education:

1. To what extent are techniques for establishing validity consistent with the *Standards for Educational and Psychological Testing* (AERA, et al., 1999)?
2. To what extent are exploratory factor and principal component analysis methods, data analysis, and reported evidence consistent with factor analytic best practices?

This chapter provides a detailed description of the systematic review methodology employed to answer the research questions, including a review of the study design, sample, search strategy, materials, procedure, and analysis.

Study Design

Both content analysis and systematic review methodologies were reviewed as potential study designs for this research. Krippendorff (2004) defines content analysis as "a research technique for making replicable and valid inferences from texts to the context of their use" (p. 18). Small meaningful units of text are derived from the manifest

content, or the exact text as written. Using clear, transparent, replicable rules, these meaning units, through an emergent design, inform relevant categories for their organization. Subsequently, through further analysis, the researcher moves from data specific categories to higher levels of abstraction that allow for meaning making of the text within its context. Although content analysis offers a systematic approach and a focus on written text, content analysis did not meet the needs of this research study. An emergent design was determined to not support the research questions where an *a priori* set of best practices needed to be extracted specifically from the medical education instrument development literature.

Therefore, to address the two research questions for this study, a systematic review was conducted, informed by the Cochrane Collaboration and the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPICC). The Cochrane Collaboration supports systematic reviews of the effects of treatment interventions in human healthcare to inform both medical practitioners and health policy leaders. The Evidence for Policy and Practice Information and Co-ordinating Centre (EPPICC) focuses more broadly on systematic reviews in the social sciences and public policy. The definition of systematic reviews espoused by the EPPICC extends the focus beyond exclusively understanding the effects of interventions, "systematic reviews aim to find as much as possible of the research relevant to the particular research questions, and use explicit methods to identify what can reliably be said on the basis of these studies" (Evidence for Policy and Practice Information and Co-ordinating Centre (EPPICC), 2010). The Cochrane Collaboration and the EPPICC have in common the articulation of three key criteria for systematic reviews: (a) a comprehensive review of research

evidence delimited by eligibility criteria, (b) explicit, transparent, reproducible methods, and (c) a systematic approach to the organization and presentation of findings from the reviewed studies (EPPICC, 2010; Green, Higgins, Alderson, Clarke, Mulrow, Oxman, 2008). Based on these three criteria, a systematic review seemed best able to provide a research design that produces comprehensive, replicable findings to answer the two research questions for this study. The following documentation presents how this research study complies with these three expectations.

Sample

All primary empirical medical education research articles that met the following criteria were eligible for inclusion in the review: (a) human study, (b) development of a new or revised instrument, (c) application of exploratory factor analysis or principal component analysis, (d) written in English, and (e) published January 2006 through December 2010. Review articles, editorials, qualitative studies, and case discussions were excluded. Principal component analysis (PCA) studies were included in order to examine how often PCA was used in place of exploratory factor analysis (EFA). Historically, systematic reviews generally cover a five- or ten-year time period. To address feasibility issues for this study, a five-year range was selected. If a study combines an EFA with a follow up confirmatory factor analysis (CFA), only the EFA methods and reported evidence were reviewed. Studies employing only CFA within instrument development were excluded. Again, the exclusion of CFA articles was determined based on practicability. If one article included more than one instrument developed using EFA or PCA, each instrument was reviewed separately. In addition, if

one instrument involved development using more than one factor analysis, each factor analysis was coded separately.

Search Strategy

A systematic approach to searching the literature was applied based on the eligibility criteria through an electronic search of MEDLINE, Educational Resources Information Centre (ERIC), PsycINFO, and Cumulative Index to Nursing and Allied Health Literature (CINAHL) databases. Variations of 10 search terms were used as they were represented in the thesaurus of each database, including *validity, reliability, test construction, factor analysis,* and *medical education*. In addition, the reference lists of all included articles were hand searched.

An electronic search conducted December 2010 using the eligibility criteria – (a) human study, (b) development of a new or revised instrument, (c) application of exploratory factor analysis or principal component analysis, (d) written in English, and (e) published between 2006 and 2010 – identified 898 potentially relevant articles. This search was across multiple databases, so these numbers likely include duplicates. Titles and abstracts were reviewed to determine inclusion or exclusion. Based on this process, 791 articles were excluded. Again, using the eligibility criteria, a full-text review of the remaining articles resulted in further exclusion of articles that did not meet the inclusion criteria. This search and review process identified 60 articles for the review.

Next, a hand search of the reference lists from the included articles identified 12 articles for inclusion. After accounting for duplicates across the electronic and hand search, a total of 62 articles were included in this systematic review (See Figure 2). The

full-text for each of these 62 articles was retrieved using electronic databases and inter-library loan provided by the Virginia Commonwealth University library system.

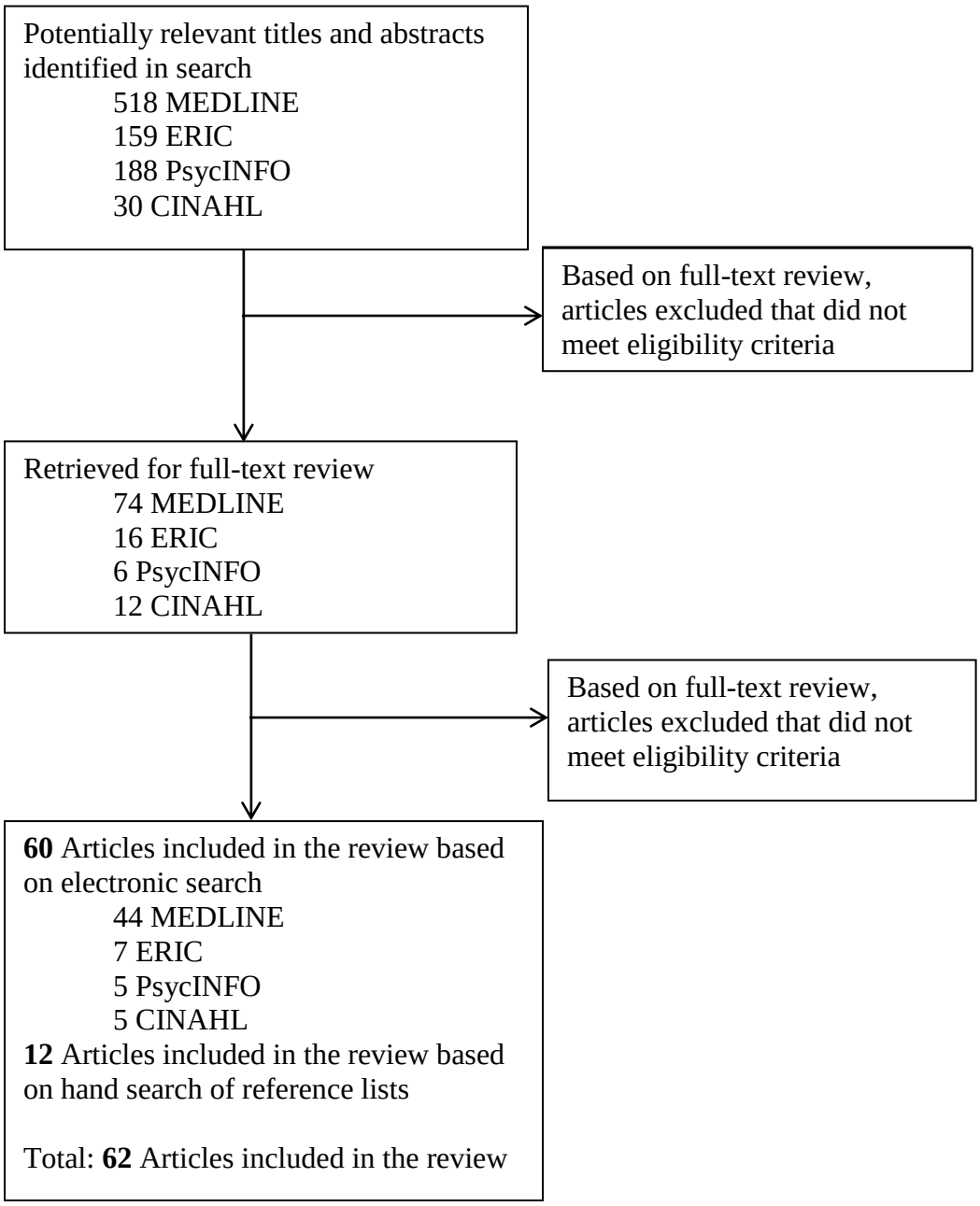

Figure 2. Search details*

*Categories may not be mutually exclusive.

Materials and Procedures

A data extraction form and coding manual, informed by the *Standards for Educational and Psychological Testing* (AERA et al., 1999) and best practices in factor analysis, were created (See Appendix A and Appendix B). The standardized data entry form and reference manual provide a systematic process for extraction of factor analysis methods and reported evidence for establishing validity from each article included in the review. Recommendations from the *Cochrane Handbook for Systematic Reviews of Interventions* (2008) were incorporated into the design of the data extraction form and manual. Related to formatting, the data extraction form includes documentation of the article title, authors, journal, year published, coder name and space for documentation of any notes by the coder. The coder documented the construct measured using an open-ended response format. For each data point, tick boxes or coded responses were used to reduce coder error and increase efficiency. The options "not reported" or "unclear" were included in addition to yes/no or other categorical response options. The *Cochrane Handbook for Systematic Reviews of Interventions* (2008) emphasizes the importance of "detailed instructions to all authors who will use the data collection form" (n.p.); thus, a coding manual was developed as a reference to provide the coders with instructions to help standardize the coding process.

Pilot Study. This structured data extraction form, including three sections, (a) educational outcome level, (b) factor analysis, (c) other techniques for establishing validity evidence, was pilot tested using select peer-reviewed instrument development articles (n = 5) published in 2005, prior to the proposed review time frame of 2006-2010. Using the same search strategy previously described, five eligible articles were retrieved

for full-text review. The researcher coded all five articles using the data extraction form, taking detailed notes of necessary revisions to the form and guide to clarify both structure and process. Revisions were made to both the data extraction form and coding manual based on the pilot study findings. An example of one revision is the refinement of the traditional validity terms and definitions. The original form and coding manual are provided for reference (See Appendix C and D).

Second Coder Training. A second individual with expertise in the content area was trained to use the revised data extraction form and coding manual. This individual is a doctoral student in the Research and Evaluation track of the doctorate of philosophy in education program within the Virginia Commonwealth University School of Education. Training involved a three step process: (a) self-study, (b) in person, hands-on coding training with sample articles, (c) independent coding and agreement calculation.

Self-study. First, the second coder was provided a hard copy of chapters one, two, and three of this dissertation, including full reference information, and a copy of the revised data extraction form and coding manual. After a two week self-study period, the second coder was provided one article (Aukes, Geertsma, Cohen-Schotanus, Zwierstra, & Slaets, 2007) selected from the pool of 62 articles included in the review, to be coded using the data extraction form and coding manual prior to the first in person training session. The lead researcher also coded this article in advance of the in person training session. Following the initial application of the coding form and manual on the Aukes and colleagues (2007) article, the second coder documented questions and comments derived from the experience of coding the first article and shared these electronically with the researcher. In response, the researcher provided clarification and updated the form

based on the second coder's comments. These communications and revisions from the iterative developmental phase of the form and manual are reported in Appendix E Section I.

In Person Training. Next, both coders met in person for a two and one half hour session. To begin, it was confirmed there were no questions about the self-study materials; therefore, the session began with a review of the coding by each coder for the first article (Aukes et al., 2007) through discussion of each section on the data extraction form to examine agreements and disagreements. Disagreements were resolved through consensus and informed further revisions to the form and manual. The form and manual were updated together during the session, and changes are documented in Appendix E Section II.

The second half of the training session involved independent coding of a second article also from the overall sample of 62 articles (Tian, Atkinson, Portnoy, & Lowitt, 2010) followed by a review of coding by each coder to evaluate agreements and disagreements. Again, this process pointed to minor revisions to the form and manual which are documented in Appendix E Section II. Overall, disagreements in coding for these two articles were minimal and easily resolved; therefore, the researcher and second coder agreed to move forward with the final phase of training, the independent coding of three articles (Frye, Sierpina, Boisaubin, & Bulik, 2006; Sargeant, Hill, & Breau, 2010; Wright, Levine, Beasley, Haidet, Gress, Caccamese, Brady, Marwaha, & Kern, 2006) to allow for an initial calculation of coder agreement.

Independent Coding for Initial Agreement Calculation. Following the in person training session, the researcher and second coder were provided a copy of three articles

(Frye et al., 2006; Sargeant et al., 2010; Wright et al., 2006) and the updated coding manual and form reflecting changes based on the first two rounds of coding from the in person training session. The researcher and coder allotted one week for the coding of the three articles. The researcher coded the articles first leading to further minor revisions to the form (See Appendix E Section III) and provided an electronic copy of the revised materials to the second coder who subsequently applied the manual and form to the three articles. Questions and comments from the second coder were documented and are reported in Appendix E Section III.

The researcher and second coder met in person for a one hour 15 minute session to review coding for the three articles. Disagreements and agreements were reviewed; disagreements were resolved by consensus and led to final revisions to the form and manual (See Appendix E Section III). To provide an initial estimate of agreement between coders, the proportion of agreement, calculated as total number of agreements divided by total number of agreements plus disagreements, was determined. Although, Cohen's Kappa provides a coefficient of agreement appropriate for categorical coding and accounts for chance agreement (Cohen, 1960), it is only well suited for dichotomous data. Weighted Cohen's Kappa is available to enable agreement calculation for ordinal coding. However, this study has variables with multiple nominal values; therefore, Cohen's Kappa or weighted Kappa did not support the present need for agreement calculation; therefore, the proportion of agreement was instead utilized. The proportion of agreements, calculated as total number of agreements divided by total number of agreements plus disagreements, was calculated for the open-ended variables such as the sample size reported in the study. With high agreement (91.2%) for these three articles,

the researcher and coder transitioned to the full data extraction phase using the final data extraction form and coding manual (See Appendix A & B).

Data Extraction. Based on the pilot study of five articles and the three-round, iterative developmental phase using an additional five articles, the final coding manual and data extraction form included four sections: (a) descriptive information about the article, (b) educational outcome level, (c) factor analysis methodological decisions and reported evidence, and (d) other techniques for establishing validity evidence. The researcher utilized the final materials to extract systematically the data from all 62 articles included in the review; the five articles coded during self-study, in person training, and independent coding were coded again using the final versions of the coding manual and form. The second coder was assigned a randomly selected 10% (n = 6) of all articles to code (Di Lillo, Ciccetti, Lo Scalzo, Taroni, & Hojat, 2009; Mihalynuk, Coombs, Rosenfeld, Scott, & Knopp, 2008; Roh, Hahm, Lee, & Suh, 2010; Singer & Carmel, 2009; Sodano & Richard, 2009; Wall, Clapham, Riquelme, Vieira, & Cartmill, 2009). Again, the researcher calculated agreement using proportion of total agreements.

First, key information for each article was documented including the title, journal, authors, volume, issue, page numbers, and publication date. In addition, to enable description of the types of instruments reviewed, the construct measured and/or the instrument title for each instrument was abstracted. The educational outcome level assessed or evaluated by the instrument was coded using the Moore et al. (2009) Outcomes Framework – level 1: participation; level 2: satisfaction; level 3A: declarative knowledge; level 3B: procedural knowledge; level 4: competence; level 5: performance; level 6: patient health; level 7: community health (See Table 1).

Next, specifics related to methodological decisions made and reported evidence for factor analysis were coded. The total sample size and/or ratio of participants to variable were coded as reported in the article; for relevant cases, a not reported and an unclear option were available. The model of analysis reportedly used also was coded: PCA, EFA, not reported, unclear. The specific extraction method was documented: Principal Component Analysis, Maximum Likelihood, Principal Axis Factoring, Generalized Least Squares, other, combination of methods, not reported, unclear. A comparison of reported model of analysis to the extraction method creates the opportunity to evaluate whether terminology was applied incorrectly (e.g., reported using an exploratory factor analysis with principal component analysis extraction). In addition, the coder indicated whether justification for the specific extraction method was reported and reflected consideration of the items' level of measurement. Type of rotation was coded as orthogonal or oblique, and the specific rotation method was recorded, if reported. For oblique rotations, the researcher determined if both the factor correlation and factor pattern matrices were reported using the following coding options: factor correlation matrix only, factor pattern/loadings only, both, unclear, none. Using a binary yes/no option, the researcher noted if justification for the rotation method based on hypothesized or theorized relationships between factors was provided in the article. Each criterion used to determine how many factors to retain was coded: previous theory, number set *a priori*, eigenvalue greater than one rule, scree test, minimum average partial, parallel analysis, minimum proportion of variance accounted for by factor, number of items per factor, conceptual interpretability/meaningfulness, not reported, unclear, other. The minimum factor loading required for an item to load on a factor for

each study was documented; if this information was not reported or other criteria were used to determine which items on load on which factors, this was noted. The total number of items in the instrument, number of factors retained, and the number of items retained for each factor was recorded. The coder indicated whether eigenvalues were reported for retained factors, whether variance explained by each factor and/or for the total factor solution was reported, and whether factor loadings for all items were reported. Finally, the coder assessed whether a confirmatory factor analysis was warranted for the study in lieu of EFA, and if so, whether justification for this design decision was articulated.

In addition to details of the factor analysis procedure, other techniques for establishing validity evidence were extracted from each article. Based on the lag in full implementation of terminology from the contemporary framework for understanding validity evidence, as espoused in the *Standards for Educational and Psychological Testing* (AERA et al., 1999), a traditional approach was used to extract and code types of validity and reliability as reported in the article. These results were mapped onto the contemporary framework of five sources of validity evidence for interpretation (See Table 2). This approach is adopted from the Ratanawongsa et al. (2008) review of evaluation methods in continuing medical education. Specifically, articles were coded to indicate whether the following types of validity and reliability were reported: face validity, content validity, expert review, test-criterion validity (including concurrent and predictive), convergent and discriminant evidence, divergent evidence, intra-rater reliability, inter-rater reliability, test-retest reliability, test-retest stability, alternative-form reliability, and internal consistency. Though face validity evidence was documented, the

contemporary framework no longer supports the use of face validity as a source of evidence. In the contemporary framework, evidence based on consequences of testing was introduced; however, the traditional framework does not account for this type of evidence. If applied in the included articles, specific techniques for establishing validity based on this source were allowed to emerge during the review. Although construct validity is a central concept to the traditional framework of the triad of validity types – content, criterion, construct, from the contemporary perspective all validity evidence supports construct validity. Therefore, for this review, ascribing a precise definition to construct validity to allow for its extraction in a consistent, reliable, and meaningful way was not feasible. Therefore, the researcher did not extract construct validity, as a stand-alone type of validity evidence, from the reviewed articles. Rather, within this framework, all other reported evidence together constitutes support for the instrument's construct validity. In addition, other techniques for establishing validity evidence were extracted: expert review, questioning test takers about process of response to items, records capturing phases of the development of a response, dimensionality (factor analysis), item analysis, differential item functioning and differential test functioning, and pilot testing.

Details of these specific data extraction points for both factor analysis techniques and other techniques for establishing validity evidence are illustrated in the data extraction form and coder manual (Appendices A and B).

Analysis

Frequency tables provide a summary of current instrument development practice in medical education presented by educational outcome level according to the Moore et

al. (2009) outcomes framework. Specifically, a series of frequency tables summarize the factor analysis methodological practices and include frequencies for each coded response by educational outcome level and in total. Sample size is reported using frequency ranges and descriptive statistics including mean, standard deviation, and range. For other techniques for establishing validity, a second table presents the frequency of use of each specific type of reliability and validity evidence defined within a traditional classification system and mapped onto the contemporary framework of validity as a unitary concept.

The researcher compared current practice to best practices based on the *Standards for Educational and Psychological Testing* (AERA et al., 1999) and for factor analysis as they derive from the literature (Comrey & Lee, 1992; Floyd & Widaman, 1995; Gorsuch, 1983; Reise, Waller, & Comrey, 2000; Tabachnick & Fidell, 2007) to answer the research questions.

Delimitations

Two critical elements of this research study design – the conceptual framework of factor analysis best practices and the eligibility criteria – may limit the study findings. First, factor analysis best practices for this study are defined based on an extensive review of the literature related to five key methodological decision points – (a) sample, (b) model of analysis, (c) extraction method, (d) rotation method, and (e) criteria for factor and item retention. However, currently, there is no commonly accepted set of best practices. Thus, this researcher has proposed one framework for interpretation of the current findings based on the best available evidence. Second, note that only published instrument development articles using factor analysis were included. Some instrument development research may employ other techniques for establishing validity without the

inclusion of factor analysis; however, the researcher did not review these articles in this study. Focusing on articles that employ factor analysis likely predisposes the researchers to report evidence for validity based on particular sources that fit the study design and research question; whereas, instrument development more generally that does not include factor analysis may reflect different techniques for establishing validity evidence. In addition, confirmatory factor analysis articles was excluded which limits the potential to comment on current practice to exploratory factor analysis in medical education instrument development.

Institutional Review Board

This study does not involve human subject research; therefore, Institutional Review Board approval was not required.

Chapter Four

Results

Sample

A total of 62 articles were included in the systematic review of techniques for validity evidence and factor analysis methods in medical education literature based on the following inclusion criteria: (a) human study, (b) development of a new or revised instrument, (c) application of exploratory factor analysis or principal component analysis, (d) written in English, and (e) published January 2006 through December 2010. Two articles included the development of two instruments with distinct constructs; whereas 60 articles discussed the development of a single instrument, resulting in a total of 64 instruments reviewed. Fourteen of the 62 articles (22.6%) conducted more than one factor analysis; each of these analyses was coded individually for a total of 95 factor analyses reviewed. Nine of these articles used two factor analyses, three sets of authors conducted three factor analyses, one study involved eight analyses, and the final article reported on 12 separate factor analyses. For the most part, these multiple factor analyses represent the inclusion of two separate samples within one study, either a pilot and testing sample or two samples from distinct sampling frames, where a factor analysis was conducted on each sample and then results were compared.

Within the five-year range (2006-2010) of instruments studied, the distribution of articles by year is rather consistent. From 2006 and 2007, 10 (16.1%) articles were included for each year. Nine (14.5%) of the reviewed articles were published in 2008, 22 (35.5%) in 2009, and 11 (17.7%) in 2010. Thirteen articles (21%) were published in *Medical Teacher*, 10 articles (16.1%) in *Academic Medicine*, and five (8.1%) in each of

two journals, *Medical Education* and *Journal of General Internal Medicine*. The remainder of the articles came from a range of publications in medical education, specialty medicine, and higher education (See Table 3).

Table 3
Distribution of reviewed articles (n = 62) by journal and year of publication

	Year of publication					
Journal	2006	2007	2008	2009	2010	Total
Medical Teacher	-	3	-	7	3	13
Academic Medicine	1	1	-	5	3	10
Medical Education	1	2	-	1	1	5
Journal of General Internal Medicine	2	2	-	1	-	5
Advances in Health Sciences Education	1	-	1	-	-	2
Education for Health	-	-	1	-	1	2
Patient Education and Counseling	-	-	1	1	-	2
Adult Education Quarterly	1	-	-	-	-	1
American Journal of Obstetrics and Gynecology	-	-	1	-	-	1
American Journal of Preventative Medicine	1	-	-	-	-	1
Anatomical Sciences Education	-	-	-	1	-	1
Annals of Academic Medicine Singapore	-	-	1	-	-	1
Archives of Pathology and Laboratory Medicine	-	-	-	1	-	1
Assessment and Evaluation in Higher Education	-	-	1	-	-	1
BMC Medical Education	1	-	-	-	-	1

BMC Medical Informatics and Decision Making	-	-	1	-	-	1
British Journal of Educational Technology	-	1	-	-	-	1
Canadian Journal of Rural Medicine	-	-	-	1	-	1
Clinics	-	-	1	-	-	1
Croatian Medical Journal	-	1	-	-	-	1
International Psychogeriatrics	-	-	-	1	-	1
Journal of Career Assessment	-	-	-	1	-	1
Journal of Continuing Education in the Health Professions	-	-	-	-	1	1
Journal of Emergency Medicine	-	-	-	-	1	1
Journal of Interprofessional Care	1	-	-	-	-	1
Journal of the American College of Nutrition	-	-	1	-	-	1
Journal of Vocational Behavior	-	-	-	1	-	1
Medical Education Online	1	-	-	-	-	1
Revista Brasileira de Anestesiologia	-	-	-	1	-	1
Teaching and Learning in Medicine	-	-	-	-	1	1
Total	10	10	9	22	11	62

Grouping instruments based on the construct measured resulted in 14 meaningful groups including measures of the following: (a) clinical content specific knowledge, skills, or attitudes (*n* = 10); (b) career preference assessments (*n* = 7); (c) professionalism (*n* = 7); (d) educational environment (*n* = 5); (e) instructional quality (*n* = 5); (f) communication and feedback skills (*n* = 5); (g) self-directed/lifelong learning (*n* = 4); (h) empathy (*n* = 4); (i) learning styles/behaviors/skills (*n* = 4); (j) interprofessional teams, teams, and team leadership (*n* = 3); (k) patient safety (*n* = 2), and (l) educational program quality (*n* = 2). The remaining six articles fall into a miscellaneous category. Four instruments were investigated in more than one study either using an adapted version of the instrument or by applying it to a new population (e.g., students instead of physicians). Specifically, the Postgraduate Hospital Educational Environment Measure (PHEEM) represented five of the 64 instruments, and the Jefferson Scale of Physician Empathy (JSPE) and the Jefferson Scale of Physician Lifelong Learning (JeffSPLL) each comprise three of the 64 instruments. Using Moore et al.'s (2009) outcomes framework, 13 (20.3%) of the instruments reviewed evaluated programs at level 2 in Moore et al.'s (2009) framework for levels of assessment and evaluation outcomes. This level 2 measures participant satisfaction. Thirty-six (56.3%) instruments assessed level 3A: declarative knowledge/attitude. Four (6.3%) instruments measured competence in an educational setting (level 4); eight instruments (12.5%) represented outcome measures of performance of residents and/or physicians in practice (level 5). For three (4.7%) instruments, it was unclear what level outcome the instrument measured according to Moore et al.'s (2009) outcomes framework; most often, this occurred because authors failed to include the specific items or to report the level of outcome measurement in the

publication. No articles reviewed for this study contained outcome measures at level 3B: procedural knowledge; level 6: patient health; or level 7: community health. Level 1: participation would not realistically be measured using an instrument; therefore, level 1 outcomes are not reflected in this review.

This study's researcher coded all 62 articles, and a trained second coder double coded a randomly selected sample of 10% (n = 6) in a peer review process. Proportion of agreements to agreements plus disagreements for the six double coded articles was 93.4% with a range from 80.9% to 100%.

Data Extraction: Techniques for Establishing Validity Evidence

Before examining individual techniques for establishing validity evidence, it is important to note that eight articles reviewed as part of this study reported reliability or validity evidence from previous empirical investigations of the instrument, yet they failed to pursue evidence for reliability and validity within the context of the current application. For example, one study provided a description of previously established evidence based on test content, including expert review; however, in the first instance, the instrument measured the construct in the general population, and the authors did not consider the relevance of the content and items in the second instance when the measurement was applied to medical students. Evidence derived from previous investigations of instruments was not reported.

Borrowing from the methodology applied by Ratanawongsa et al. (2008), the researcher extracted techniques for establishing validity evidence from the reviewed articles using the traditional validity framework (e.g., content validity, construct validity, criterion validity). These terms were then mapped onto the contemporary validity

framework supportive of validity as a unitary concept with multiple sources of supporting evidence (See Table 2). Authors often utilized the term "construct validity," yet from a contemporary perspective, all validity evidence is evidence of construct validity. Therefore, this was not specifically addressed in the review as a stand-alone technique.

Evidence Based on Test Content. Overall, 23 (35.9%) of the 64 reviewed instruments were supported by one source of evidence based on test content (e.g., traditional content validity, expert review, or pilot test); 17 (26.6%) were supported by two sources of evidence, and nine (14.1%) instruments were accompanied by three sources of evidence based on test content. For forty-four (68.6%) of the instruments, the authors reported evidence coded using the traditionally understood meaning of content validity. For example, 25 of the 44 instruments included items developed based on a review of the literature, or based on key competencies or core content as defined by a national agency or organization affiliated with the measured construct. A sample from the target population reviewed sixteen of the 45 instruments for content and clarity through a focus group discussion or pretest. Moreover, for the newly developed instruments, nine included items from previously tested questionnaires and assessments. Further, authors employed expert review of items for 24 (37.5%) of the total 64 instruments; however, the qualifications of the reviewers as experts were not always made clear. Pilot testing with the target population occurred for twenty-five percent (n = 16) of the instruments. The sample size for the pilot studies ranged from three to 878 (m = 148.67, sd = 258.85); authors failed to report the sample size in four pilot studies. Although a term affiliated exclusively with the traditional validity framework and no longer supported in the contemporary understanding of validity evidence, in the

investigation of 11 (17.2%) of the 64 instruments, authors reported face validity as support for content validity. Table four describes details of each study as reported by outcome level (Moore et al., 2009).

Table 4
Reported evidence for reliability and validity in medical education instrument development articles employing factor analysis abstracted using a traditional validity framework and mapped to the contemporary framework of validity as a unitary concept

Validity evidence	Level 2: Satisfaction *n* = 13	Level 3A: Declarative knowledge *n* = 36	Level 4: Competence *n* = 4	Level 5: Performance *n* = 8	Unclear *n* = 3	Total *n* = 64
Evidence based on test content						
Face validity	1	7	-	3	-	11 (17.2)
Content validity	9	22	4	7	2	44 (68.6)
Expert review	-	16	2	4	2	24 (37.5)
Pilot test	2	9	2	2	1	16 (25)
Evidence based on relationships with other variables						
Concurrent criterion validity	-	4	1	1	-	6 (9.4)
Predictive criterion validity	-	-	-	-	-	-
Convergent evidence	1	4	1	2	-	8 (12.5)

Discriminant evidence	-	1	-	-	-	1 (1.6)
Divergent evidence	5	16	1	2	1	25 (39.1)
Evidence based on response process						
Intra-rater reliability *Potential n* = 6	-	-	-	-	-	-
Inter-rater reliability *Potential n* = 6	-	1	1	1	-	3 (50)
Test-retest Reliability	2	2	-	-	-	4 (6.3)
Test-retest Stability	1	3	-	-	-	4 (6.3)
Questioning test takers about process of response to items	1	3	-	1	-	5 (7.8)

(e.g., cognitive interviewing)						
Evidence based on internal structure						
Internal consistency	11	35	3	8	2	59 (92.2)
Alternative-form reliability	-	-	-	-	-	-
Other techniques						
Item analysis	1	7	-	3	-	11 (17.2)
Back language translation by expert	2	4	1	-	-	7 (10.9)
Generalizability theory	1	-	2	1	-	4 (6.3)
Feasibility analysis	1	-	2	2	-	5 (7.8)
Rand coefficient	-	1	-	-	-	1 (1.6)
Tucker's phi coefficient	-	-	-	-	1	1 (1.6)

Source: AERA et al., 1999; Moore et al., 2009; Nunnally & Bernstein, 1994; Ratanawongsa et al., 2008; Trochim, 2006

Evidence Based on Relationships with Other Variables. The following five traditional validity terms relate to the contemporary validity concept of evidence based on relationships with other variables: (a) concurrent criterion validity (i.e., degree to which an instrument produces the same results as another accepted, validated, or even "gold standard" instrument that measures the same construct), (b) predictive criterion validity (i.e., degree to which a measure accurately predicts something it should theoretically be able to predict), (c) convergent validity (i.e., degree of agreement between measurements of the same construct obtained by different methodologies), (d) discriminant validity (i.e., degree to which a measure produces results different from the results of another measure of a theoretically unrelated construct), and (e) divergent validity (i.e., ability of a measure to yield different mean values between relevant groups). Convergent evidence accompanied eight (12.5%) instruments. For example, in a study investigating correlations with scores on a physician lifelong learning instrument, Hojat and colleagues (2009) correlated self-reported number of publications with the number of publications extracted from electronic databases to provide convergent evidence. Authors investigated concurrent criterion evidence for six (9.4%) instruments, and discriminant evidence was reported for only one (1.6%) instrument. For example, in one study, authors correlated scores on a newly developed measure of personal growth in residents with scores from the Ryff's validated measure of personal growth (Wright, Levine, Beasley, Haidet, Gress, Caccamese, Brady, Marwaha, & Kern, 2006). Haidet et al. (2008) examined both concurrent criterion and discriminant evidence of the CONNECT instrument, an instrument designed to measure both physician and patient explanatory models of illness, through testing of hypothesized relationships between scores on the

CONNECT subscales and previously validated instruments. Specifically for discriminant evidence, the authors examined correlations between scores on the CONNECT subscale labeled "meaning" and the well validated SF-12 instrument's physical function subscale score, expecting to find a negative correlation, asserting that "an illness with greater meaning would correlate with lower physical functioning scores" (Haidet, O'Malley, Sharf, Gladney, Greisinger, & Street, 2008, p.234). Predictive criterion evidence did not appear in the evidence for any of the 64 instruments. Divergent validity evidence was reported for 25 instruments (39.1%). For example, Hojat and colleagues (2009) examined differences between full-time clinicians and academic clinicians on orientation toward lifelong learning scale scores.

Evidence Based on Response Process. The *Standards for Educational and Psychological Testing* (1999) endorse the following as valuable ways to understand the response process and its relationship to the measured construct: (a) observations of participants in performance based outcome measures, (b) records documenting phases of the development of a written response, or (c) results from questioning participants about their response to particular items either during or after administration of the instrument. However, since the 64 instruments reviewed in this study all include numeric, closed-ended response options, the opportunity for application of the first two techniques is not available as it would be for observations, essays or other open-ended responses. A similar mechanism to understand the response process of respondents is to question them about the process of response either during administration of the instrument or immediately following (e.g., cognitive interviewing). This is different from asking a sample from the target population to comment on the thoroughness or clarity of items;

rather, this specifically asks respondents to discuss the process of response (e.g., how they interpret the language of the item, how they understand the response options, how they select a response option). Of the 64 instruments, this method was used for five (7.8%). Authors sought evidence of stability over time for eight instruments: specifically, test-retest reliability around a two week interval for four (6.3%) instruments and test-retest stability around a six month interval for four (6.3%) instruments. One final source of evidence based on response process comes from inter-rater and intra-rater reliability; yet, this source is only relevant to instruments that involve multiple raters evaluating the same construct for the same evaluand (e.g., medical student, resident, or physician) or individual raters evaluating the same construct across multiple evaluands. Of the 64 instruments in this review, only six instruments included either multiple raters evaluating the same construct for the same evaluand or individual raters evaluating the same construct across multiple evaluands; therefore, this source of evidence was relevant to only these six instruments. Of the six, three (50%) reported inter-rater reliability, but none reported intra-rater reliability. Table four lists the details of evidence based on response process by outcome level.

Evidence Based on Internal Structure. As this review was limited to studies that employed factor analysis, reporting for all 64 instruments included evidence based on dimensionality to support internal structure. However, the empirical evidence to support dimensionality was not always linked back to theoretical evidence for a uni- or multi-dimensional construct. Authors reported evidence for internal consistency for almost all (n = 59, 92.2%) of the instruments reviewed. Although internal consistency was most often estimated from Cronbach's alpha, item-scale and item-total correlations

and reliability-if-item-deleted also were applied and, in turn, used to determine which items to retain based on their contribution to the instrument's dimensionality and reliability. Alternative-form reliability was not used as supporting evidence for any of the 64 instruments.

Evidence Based on Consequences of Testing. Evidence based on consequences of testing might include clear description of the process of scoring, reporting of cut-off scores applied and justification of these scores, calculation and reporting of classification accuracy when relevant, and reporting of the standard error measurement (AERA et al., 1999; Downing, 2003). Further, examination of outcomes caused by the assessment - positive and negative, as well as intended and unintended - would relate to this source of evidence (Andreatta & Gruppen, 2009; AERA et al., 1999). For the 64 instruments reviewed, the authors did not report evidence based on consequences of testing.

Other Techniques for Establishing Validity Evidence. This review identified a number of additional techniques applied in these studies that are associated with quality instrument development that can lead to further reliability and validity evidence. The researcher identified analysis of the individual items applied in eleven (17.2%) of the 64 instruments, including examination of variability in response and patterns of non-response. This analysis led to the deletion of some items that lacked variability and those items whose patterns of non-response suggested problems with the item language or item content. Seven (10.9%) studies that involved the adaptation of an existing instrument to a new language employed the use of back language translation by language experts in the original and translated languages. This involved first translating the original instrument into the new language. Then, an expert translated it back into the original language, and

finally, a comparison was made by a language expert between the version translated back into the original language and the original instrument to ensure consistency in meaning. Authors conducted generalizability theory analysis for four (6.3%) instruments to determine the number of raters or the number of times the evaluand would need to be evaluated. Authors conducted feasibility analysis for five (7.8%) instruments, which included surveying or discussion with respondents on the feasibility of completing the instrument concerning factors such as time to complete or accessibility of the instrument. Finally, the Rand coefficient and Tucker's phi coefficient were each reported for a single instrument. In the one study, the Rand coefficient compared the empirically-derived factor structure to the theoretically based structure proposed by experts in the topic; the coefficient ranges from 0 to 1 and a coefficient of 0.89 were reported (Short, Alpert, Harris, & Surprenant, 2006). Tucker's phi coefficient provides a correlation between the factors derived from two independent samples. In this instance, Tromp and colleagues (2010) used this approach to estimate congruence of the two-factor solution between general practitioner trainers and general practitioner trainees on a measure of professionalism. Table four presents these other techniques by outcome level.

Overall, since this review only included articles that conducted factor analysis, when dimensionality as a source of validity evidence was excluded, 59 (92.2%) of the 64 instruments were supported by at least one source of both reliability and validity evidence. Only validity evidence was reported for the remaining five instruments.

Data Extraction: Factor Analysis Methods

Sample Size. The sample size utilized for factor analysis ranges across the 95 analyses from a low of 45 to a high of 91,073. The mean was 1386.17 (*sd* = 9737.28);

however, this distribution is positively skewed. By removing the single 91,073 sample outlier, the mean is reduced to 343.3 (*sd* = 444.45). The median sample size for the 95 factor analyses reviewed was 208. Specifically, 13 (13.7%) factor analyses were run on a sample size of less than 100. Twenty-five (26.3%) of the factor analyses used sample sizes of between 101 and 200. Twenty-four (25.3%) of the analyses were conducted with between 201 and 300 respondents. Sample sizes ranging from 301 to 400 respondents were employed in nine (9.5%) analyses; samples of 401 to 500 were reported in three (3.2%) studies; and sample sizes greater than 500 represent 13 of the analyses (13.7%). For the remaining eight (8.4%) factor analyses, the sample size was unclear. Of the 87 factor analyses that reported sample size, 83 also reported the total number of items in the final instrument, allowing for calculation of the participant to item ratio. This value ranges from 1.54 participants per single item (1.54:1) to 3140.45 participants per single item (3140.45:1) (or 115.14:1 if removing the largest sample as an outlier); the mean is 55.7 participants per single item (55.7:1), and the median is 11.55 participants per single item (11.55:1). Table five reports frequency by outcome level assessed or evaluated based on Moore et al.'s (2009) outcomes framework.

Table 5
Sample size as reported in medical education instrument development articles employing factor analysis (n = 95)

Sample size	Level 2: Satisfaction *n* = 15	Level 3A: Declarative knowledge *n* = 40	Level 4: Competence *n* = 4	Level 5: Performance *n* = 20	Unclear *n* = 16	Total *n* = 95
100 and below	5	3	1	1	3	13 (13.7)
101-200	3	11	1	8	2	25 (26.3)
201-300	-	10	1	3	10	24 (25.3)
301-400	3	4	1	-	1	9 (9.5)
401-500	-	3	-	-	-	3 (3.2)
501 and above	3	7	-	3	-	13 (13.7)
Unclear	1	2	-	5	-	8 (8.4)

Source: Moore et al., 2009

Model of Analysis and Extraction Method. Of the 95 factor analyses reviewed across 62 articles, principal component analysis as a model and extraction method was most frequently applied in these studies (n = 60; 63.2%). In comparison, 16 (16.8%) factor analyses employed a common factor or exploratory factor model. However, thirty-five of the 95 analyses were termed exploratory factor analyses by the authors, yet 18 (18.9%) were, in fact, principal component analyses. In addition, three articles incorrectly reported the utilization of a confirmatory factor analysis model when an exploratory factor analysis was applied to assess consistency between the factor solution and a hypothesized, theoretical, or previous empirically defined factor structure. Of those analyses based on the common factor model, 5 (5.3%) employed principal axis factoring as the extraction method, eight (8.4%) utilized maximum likelihood extraction, two (2.1%) used unweighted least squares, and one (1.1%) used weighted least squares. In three analyses (3.2), the extraction method was unclear. Overall, for 16 (16.8%) of the 95 factor analyses, the extraction method was not reported. In addition, only one (1.1%) analysis in the review provided justification for the extraction method based on consideration of the level of measurement of the items. See table six for complete details for extraction method by outcome level.

Table 6
Extraction method as reported in medical education instrument development articles employing factor analysis (n = 95)

Extraction method	Level 2: Satisfaction *n* = 15	Level 3A: Declarative knowledge *n* = 40	Level 4: Competence *n* = 4	Level 5: Performance *n* = 20	Unclear *n* = 16	Total *n* = 95
Principal components analysis (PCA)	6	21	2	15	16	60 (63.2)
Common factor Model						
Principal axis factoring (PAF)	-	4	-	-	-	4 (4.2)
Maximum likelihood	3	4	-	1	-	8 (8.4)
Weighted least squares	-	-	-	1	-	1 (1.1)
Unweighted least squares	-	1	1	-	-	2 (2.1)
Combination: PCA	-	1	-	-	-	1 (1.1)

and PAF*						
Unclear	-	2	1	-	-	3 (3.2)
Not reported	6	7	-	3	-	16 (16.8)

*In

*For this instance, both PCA and PAF extraction methods were applied; the PAF solution was interpreted.
Source: Moore et al., 2009

Rotation Method. Regarding factor rotation methods, seven (7.4%) of the factor analyses applied a combination of orthogonal and oblique factor rotations; of these seven, all interpreted the orthogonal rotation. Overall, 62 (65.3%) of the 95 factor analyses interpreted an orthogonal rotation. Specifically, 61 (64.2%) utilized a varimax rotation, and one rotation was described only as an orthogonal rotation with no specificity of the rotation type. A smaller percentage of studies (n = 20; 21.1%) interpreted an oblique rotation. Overall, for oblique rotations, seven (7.4%) were promax, 17 (17.9%) were direct oblimin, and two failed to articulate the exact oblique rotation type. Both factor pattern matrices (i.e., factor loadings) and factor correlation matrices (i.e., correlations between factors) should be reported for oblique rotations to aid in interpretation. Of the 20 oblique rotations in this review, 12 (60%) did report both factor pattern and factor correlation matrices, two (10%) reported only factor correlations, two (10%) reported only factor loadings, and four (20%) reported neither. For ten (10.5%) of the 95 factor analyses, the factor rotation was not reported, and for two (2.1%) it was unclear. Justification for the selection of a specific rotation method based on theoretical or empirical evidence for the relationships between factors was provided for only 25 (26.3%) of analyses. In fact, three studies provided evidence for moderate to strong (>.32) correlations between the empirically derived factors, yet interpreted the orthogonal rotation in error. Table seven provides frequencies in total and by outcome level for further detail.

Table 7
Rotation method as reported in medical education instrument development articles employing factor analysis (n = 95)

Rotation method	Level 2: Satisfaction *n* = 15	Level 3A: Declarative knowledge *n* = 40	Level 4: Competence *n* = 4	Level 5: Performance *n* = 20	Unclear *n* = 16	Total *n* = 95
Orthogonal	12	18	2	15	8	55 (58)
Varimax	12	23	2	16	8	61 (64.2)
Not reported	-	1	-	-	-	1 (1.1)
Oblique	1	9	-	2	9	20 (21.1)
Promax	1	4	-	2	-	7 (7.4)
Direct oblimin	-	8	-	1	8	17 (17.9)
Not reported	-	2	-	-	-	2 (2.1)
Unclear	-	1	-	-	-	1 (1.1)
If oblique, which coefficients were reported? *n* = 20						
Factor correlation only	-	2	-	-	-	2 (10)
Factor pattern only	1	-	-	1	-	2 (10)
Both	-	4	-	-	8	12 (60)
None	-	3	-	1	-	4 (20)

Combination orthogonal and oblique	-	6	-	1	-	7 (7.4)
No rotation	-	-	1	-	-	1 (1.1)
Not reported	2	5	1	2	-	10 (10.5)
Unclear	-	2	-	-	-	2 (2.1)

Source: Moore et al., 2009

Criteria for Factor Retention. Overall, 42 (44.2%) of the factor analyses applied only one criterion in determining the number of factors to retain. Thirty (31.6%) reported using two criteria, and 12 (12.6%) considered three or more criteria in selecting which factors to retain in the solution. Similar to reporting of the rotation method, the remaining 11 (11.6%) articles failed to report which criteria were used. In particular, the Kaiser criterion, or eigenvalue greater than one rule, and the Cattell scree test were most commonly applied. The Kaiser criterion was used in 46 (48.4%) factor analyses, and the Cattell scree test in 35 (33.7%). Twenty-one (22.1%) of the analyses considered the conceptual interpretability or meaningfulness of each factor when making decisions on which factors to retain, and 18 (19%) set a minimum number of items required per factor for retention. Other methods were used less frequently. These include: (a) a minimum proportion of variance accounted for in the factor solution (n =5, 5.3%), (b) previous theory as a guide to the number of factors to retain (n = 4, 4.2%), (c) parallel analysis (n = 4, 4.2%), (d) χ^2 statistic within maximum likelihood extraction (n = 3, 3.2%), and (e) a number of factors set *a priori* (n = 2, 2.1%). The Cattell-Nelson-Gorsuch objective scree test; minimum average partial; Mokken scale analysis, an established minimum internal consistency per scale; and simple structure were individual criterion each applied one time (1.1%) in the 95 analyses. Further details on criteria for factor retention can be seen in table eight.

Table 8
Criteria used to determine the number of factors to retain as reported in medical education instrument development articles employing factor analysis (n = 95)

Criteria for factor retention	Level 2: Satisfaction *n* = 15	Level 3A: Declarative knowledge *n* = 40	Level 4: Competence *n* = 4	Level 5: Performance *n* = 20	Unclear *n* = 16	Total *n* = 95
Previous theory	-	4	-	-	-	4 (4.2)
A priori	-	2	-	-	-	2 (2.1)
Kaiser criterion: eigenvalue greater than 1 rule	10	20	3	10	3	46 (48.4)
Cattell scree test	5	18	2	10	-	35 (33.7)
Cattell-Nelson-Gorsuch objective scree	-	1	-	-	-	1 (1.1)
Minimum average partial	-	1	-	-	-	1 (1.1)
Parallel analysis	-	3	1	-	-	4 (4.2)
Minimum proportion of variance accounted for in solution	1	4	-	-	-	5 (5.3)
Minimum number of	2	4	-	-	12	18 (19)

items per factor						
Conceptual interpretability/ meaningfulness	-	10	-	11	0	21 (22.1)
Chi-square statistic	-	3	-	-	-	3 (3.2)
Mokken scale analysis	1	-	-	-	-	1 (1.1)
Simple structure	1	-	-	-	-	1 (1.1)
Minimum internal consistency per scale	-	1	-	-	-	1 (1.1)
Not reported	2	7	1	-	1	11 (11.6)

Source: Moore et al., 2009

Other Factor Analysis Reporting Details.

Factor Loadings. Of the 95 factor analyses, 33 (34.7%) presented a matrix including all factor loadings for all items. Thirty (31.6%) reported only factor loadings for items that met a certain loading criterion (e.g., a minimum loading value (>.40), values >.40 and <.60 on only one factor, or only the highest loading for each item). Yet, 32 (33.7%) analyses out of all 95 reported no factor loadings to communicate to the reader the details of the distribution of items across factors. Further, in a few analysis (*n* = 4), items did not meet the established criterion, yet they were not removed from the instrument, nor did the authors provide further explanation or guidance for future investigation with and use of the instrument. Almost half (*n* = 44, 46.2%) of the articles did not report the minimum factor loading required for an item to be designated as loading on a specific factor. Of the 51 (53.7%) that did report the minimum, most used a threshold of 0.40 (*n* = 32, 62.8%). Other minimum loadings ranged from 0.25 to 0.60, specifically 0.25 (*n* = 1, 2%), 0.30 (*n* = 9, 17.7%), 0.32 (*n* = 1, 2%), 0.45 (*n* = 2, 3.9%), 0.50 (*n* = 3, 5.9%), and .60 (*n* = 3, 5.9%) Table nine illustrates these findings related to factor loadings.

Factor Eigenvalues and Percentage of Variance Explained. Less than half (40%) of all studies reported the eigenvalues for each retained factor. Similar results were seen for reporting of the percentage of variance explained by retained factor (*n* = 46, 48.4%) and by factor solution (*n* = 48, 50.5%). Table nine also includes the specific distribution of reporting of eigenvalues and percentage of variance explained by outcome level. In seven of the analyses, the authors confused terminology from distinct models of analysis and stated that they conducted an exploratory factor analysis, but they reported

on the total variance explained or stated that they used a principal components model, and, subsequently, reported shared/common variance.

Confirmatory Factor Analysis versus Exploratory Factor Analysis. Finally, data were extracted to determine if a confirmatory factor analysis (CFA) would have been more appropriate in lieu of the employed model. Most factor analyses were conducted for new measures (n = 64, 67.4%); therefore, a CFA was not warranted. In addition, 24 (25.3%) of the analyses were conducted on measures that were substantially revised or tested in a new population. Again, CFA would not have been appropriate. For one study, the measure had been previously tested, but prior results failed to offer sufficient validity evidence to warrant a CFA; rather, further testing through EFA was the better solution. Three analyses (3.2%) did not require a CFA but incorporated both an EFA and CFA into the research design. In total, only three factor analyses warranted a CFA model given prior research on the instrument. One study (1.1%) did in fact conduct both an EFA and CFA; however, only two studies (2.1%) out of the 95 failed to conduct a CFA when it would have been most appropriate. Table nine provides further details on the use of CFA versus EFA.

Table 9
Other reporting details in medical education instrument development articles employing factor analysis (n = 95)

Other reporting details	Level 2: Satisfaction *n* = 15	Level 3A: Declarative knowledge *n* = 40	Level 4: Competence *n* = 4	Level 5: Performance *n* = 20	Unclear *n* = 16	Total *n* = 95
Which factor loadings were reported?						
All factor loadings for all Items	3	13	3	2	12	33 (34.7)
Limited loadings	5	12	-	13	-	30 (31.6)
None	7	14	1	5	4	31 (32.6)
Were eigenvalues reported for each retained factor?	6	20	1	11	-	38 (40)
Percentage of variance explained						
Reported by factor	7	19	2	15	3	46 (48.4)
Reported by solution	10	26	3	8	1	48 (50.5)
Was a CFA warranted?						
Yes, this was not a new measure of a new population.	1	1	-	-	-	2 (2.1)

Yes, but both EFA and CFA were done in the study.	-	1	-	-	-	1 (1.1)
No, this measure was a newly developed measure.	7	19	3	20	15	64 (67.4)
No, this measure was substantially revised or tested in a new population.	6	16	1	0	1	24 (25.3)
No, but EFA and CFA were done in the study.	1	2	-	-	-	3 (3.2)
No, the measure had been previously tested but did not offer sufficient validity evidence to warrant CFA.	0	1	-	-	-	1 (1.1)

Source: Moore et al., 2009

Chapter Five

Discussion, Recommendations, and Conclusions

Summary

The goal of this research was to address two research questions: within medical education instrument development literature, including undergraduate, graduate, and continuing medical education: (a) to what extent are techniques for establishing test validity consistent with the *Standards for Educational and Psychological Testing* (AERA, et al., 1999), and (b) to what extent are exploratory factor and principal component analysis methods, data analysis, and reported evidence consistent with factor analytic best practices? Using systematic review methodologies, a detailed review and abstraction of data from medical education instrument development studies, specifically articles employing exploratory factor or principal component analysis published in 2006-2010 (n = 62) provided results to enable the researcher to address the research questions.

Overall, for research question one, findings indicate a tendency to report validity evidence based on a specific few sources of evidence – evidence based on test content and evidence based on internal structure – with exclusion of investigation of other evidence including that based on response process, relationships with other variables, and consequences of testing. Specifically, most studies provided, in the traditional sense, at least one source of evidence based on test content. Given the eligibility criteria for inclusion in this review, it is not a surprise that all instruments included an examination of dimensionality using factor analysis. Further, almost all reported internal consistency for the subscales and total instrument, and thus provided evidence for validity based on internal structure. However, evidence based on response process and relationships with

other variables was reported less often, and evidence based on consequences of testing was not identified in this review.

Findings related to research question two are discouraging for medical education research and suggest common errors in selecting factor analysis methods and reporting evidence. Principal component analysis was dramatically overused in lieu of exploratory factor analysis even when the goal of the study was to examine dimensionality or to develop a generalizable instrument rather than data reduction. In addition, orthogonal rotations were predominantly applied and without justification despite instances of theoretical and empirical evidence to suggest an oblique rotation to be more appropriate. Nearly half of the authors mistakenly relied on only one criterion to determine the number of factors to retain in a solution. Finally, critical omissions in reporting of information were identified, such as the extraction method, rotation method, factor loadings, and minimum loading criteria, limiting the potential for replication and verification by other researchers and the evaluation by potential educators who may seek to apply the instrument in their practice.

Discussion

Validity Evidence. The body of literature reviewed in this study provides evidence of the retention of the traditional validity framework. For instance, a number of authors suggested they established the construct validity of the instrument, in the traditional sense of three types of validity – content, criterion, and construct. However, from the contemporary perspective, all validity evidence supports construct validity; therefore, this term did not always convey substantial meaning in communicating what techniques for establishing validity were applied. Only a very few studies reported

validity evidence using contemporary validity terms such as evidence based on internal structure or evidence based on test content. It is not fully clear why the transition from the traditional validity framework to the contemporary validity framework has yet to occur in medical education, despite its ten year presence. However, existing literature and resources on instrument development also retain traditional terminology that perhaps perpetuates the tradition.

All instrumentation should include supportive evidence based on test content including a detailed blueprint of the content based on a few potential sources (e.g., literature review, focus groups with participants, or expert input); expert review of the items; and pilot testing of the instrument with a sample from the target population. Although most instruments included some evidence based on test content, less than 15% of all reviewed instruments included all three of these critical elements. In addition, where expert review was employed in one-third of the studies, often the qualifications of the experts and process of review were not fully described. Pilot testing can present feasibility challenges to some research studies, particularly where access to the sample is limited. However, to the extent possible, pilot testing or at least review of potential items by a subset of the target population (which did occur more often than pilot study in this review) is highly preferred to ensure clarity and relevance of the items for the given sample.

Cognitive interviewing refers to the process of questioning respondents about the process of response either during the administration of an instrument or immediately following. Findings from this approach indicate how respondents receive, understand, and respond to the questions and should highlight any ambiguous items or response

options to help the researcher ensure that questions are eliciting the desired response. Although this method can be resource intensive, it, like pilot testing or interviews and focus groups with the target population, is a potential source of information to help refine the items of a newly developed instrument. Authors of the reviewed instruments rarely used this mechanism. An explanation for the lack of use is unclear, though it was perhaps due to resource restrictions or perhaps for some, the authors viewed the focus groups or interviews they conducted as sufficient. Another possibility for the lack of reported use of some of the techniques relates to editorial word count limits in medical education; generally, medical education journals tend to be shorter in length, which may limit what is reported in the published text.

As expected from a review limited to factor analysis studies, all of the instruments in the current review included dimensionality evidence. However, conducting an exploratory factor analysis is not, on its own, sufficient to establish evidence based on internal structure. The researcher must help establish, for the reader, the link between the empirically derived factor structure and the structure of the construct informed by the literature and previous empirical investigations. This second step was not always included in the reviewed studies, making it difficult to translate what the EFA added as supportive evidence, if anything.

Similarly, strong instrument development includes reporting of internal consistency, and almost all of the reviewed instruments included this piece of evidence for both the subscales and total scale. Cronbach's alpha was most often utilized as the internal consistency reliability statistic; yet, it is not necessarily appropriate in all internal consistency calculations. Specifically, summation of total scores is not appropriate for

multidimensional instruments; therefore, Cronbach's alpha should be limited to the subscale. McDonald (1999) purports 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. The use of omega was not identified in this review and remains unavailable in common social science statistical software programs.

Individual measures of reliability each rule out threats based on specific sources (e.g., time, multiple ratings, alternate forms). However, the reporting of multiple reliability measures together best supports the argument for reliability of an instrument. Further, generalizability theory applies a random ANOVA model to test the influence of multiple factors on reliability of an instrument. Although applied in a handful of studies in this review, this method is not generally accessible to most researchers, and the statistical assumptions often are not met in social science data limiting its applicability across studies. Test-retest reliability and stability are, however, accessible. Yet, authors failed to design these instrument development studies to enable this aspect of data collection. Although additional planning is required to accommodate stability calculations in a research design, most educational scenarios across the continuum of medical education should provide this opportunity. Medical students and residents are often highly accessible as active participants in an ongoing educational program. Within continuing medical education, contact information such as email and physical addresses are available. However, in measures of level 5, performance in practice, where patients provide feedback on physician performance, identifying opportunities for this source of reliability evidence is challenging. Multiple versions of an individual instrument were

not identified in this review making alternate forms reliability irrelevant. Approximately 10% of the instruments reviewed did include either multiple raters for an individual or a single rater who rated multiple individuals, but inter-rater and intra-rater reliability were not always reported.

Authors reported evidence based on relationships with other variables for few instruments within this review. Specifically, though divergent validity supported roughly 40% of the instruments, most did not have supporting criterion, discriminant, and convergent evidence. This is unfortunate; evidence based on relationships with other variables allows for the development of a stronger overall argument for the validity of inferences made from an instrument. The relationship between the measure and a theoretically related or unrelated measure, the demonstration of the ability of the measure to predict relevant performance, or evidence of group differences in scores on the measure based on previous theory provides important support for the proposed inferences. Evidence for validity based on relationships with other variables is only as strong as the reliability and validity of the associated variables. Therefore, perhaps appropriate measures, with rigorous reliability and validity testing, were not available for the researchers to apply in investigation of validity based on this source.

One should note almost all instruments in this review were new or substantially revised from their original versions. This implies the first step in establishing evidence for validity would include work on the content of the instrument, its structure and its relationship to the theoretical foundation. It is possible that authors are currently conducting further research with these instruments to identify evidence based on relationships with other variables or based on consequences of testing; however, this

cannot be commented upon 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 using an instrument across relevant samples and contexts.

A direct comparison of this review with previous reviews is difficult as each focused on a distinct construct and most were not oriented exclusively toward instrument development studies. Findings are variable across previous reviews, though the consensus indicates limited reporting of reliability and validity evidence (Beckman et al., 2004; Hutchinson et al., 2002; Jha et al., 2007; Ratanawongsa et al., 2008; Shaneyfelt et al., 2006; Tian et al., 2007; Veloski et al., 2005). In fact, Tian and colleagues (2007) found none of the newly developed instruments were supported by either reliability or validity evidence, and Shaneyfelt and colleagues (2006) found only 16% of studies included both reliability and validity evidence. However, almost all instruments in this review reported evidence using at least one reliability and one validity technique. Previous reviews indicate a tendency to report reliability statistics (e.g., internal consistency, test-retest reliability, or inter-rater reliability) and to employ expert review of test content (Beckman et al., 2004; Hutchinson et al., 2002; Ratanawongsa et al., 2008; Veloski et al., 2005), whereas, Shaneyfelt and colleagues (2006) found authors most often reported evidence based on relationships with other variables, followed by evidence based on test content and internal structure. Findings from the current review indicate authors most often employ techniques to support evidence based on internal structure (e.g., internal consistency).

Although more than half of all articles reported use of at least one source of evidence based on test content (e.g., expert review), one cannot conclude this evidence is complete since most articles did not report on multiple sources. These included: (a) content informed by theory and literature, (b) expert review, and (c) pilot testing. Evidence based on response process and relationships with other variables was largely underrepresented in this review, and evidence based on consequences of testing was completely absent.

Factor Analysis. Principal component analysis was the predominant model of analysis and extraction method applied in two-thirds of the reviewed analyses, despite clear statements in the literature that PCA is not appropriate for instrument development. PCA tends to inflate factor loadings, underestimate correlations between factors, and retain error in the model. This limits the potential for the factor structure to be replicated in other samples or confirmed through a confirmatory factor analysis. Further, for nearly 20% of the analyses in this study, the extraction method was unclear or not reported. Only 16% of the studies appropriately employed an exploratory factor analysis using a common factor extraction method. Overall, only one article appropriately reported justification for the selected extraction method based on the item level of measurement as recommended in the literature (Costello & Osborne, 2005; Dumenci & Achenbach, 2008; Floyd & Widamen, 1995; Muthen & Muthen, 2010; Norris & Lecavalier, 2010; Tabachnick & Fidell, 2007). These findings are consistent with previous reviews of factor analysis in psychology and general education where PCA was also most often applied (Fabrigar et al., 1999; Henson et al., 2004; Henson & Roberts, 2006). It should be noted, however, that a number of authors tangled vocabulary terms and reported they

conducted an exploratory factor analysis when a principal component analysis was actually used; this can confuse the reader and limits potential replication. These two models are not interchangeable, when data are less than ideal with low saturations or low factor loadings, PCA and EFA lead to distinctly different results that can impact the application of instrumentation in research and practice.

Similar to previous reviews (Fabrigar et al., 1999; Henson & Roberts, 2006; Pohlmann, 2004), findings from this study indicate orthogonal rotations, specifically varimax rotations, were most often applied. Oblique rotations were selected for roughly one-fifth of the studies. For approximately 10% of the analyses, the authors failed to report or failed to make clear the rotation method, and a handful reported use of an orthogonal or oblique rotation but did not specify the exact rotation method. Selection of a rotation method should derive from previous theoretical or empirical evidence that may suggest whether the researcher should anticipate correlations between factors. When evidence suggests correlated factors, an oblique rotation allows factors to correlate. On the other hand, an orthogonal rotation restricts factors, not allowing them to correlate with each other, when theoretical and empirical evidence suggests this to be appropriate. General guidance in the social sciences literature suggests an oblique rotation is always preferred to an orthogonal rotation at first, based on the assumed correlations within socio-psychological constructs. If the oblique rotation suggests correlations between factors, the researcher has additional information to aid in interpretation of the solution that might not otherwise be available through an orthogonal rotation. On the other hand, if evidence suggests that factors are, in fact, unrelated, an orthogonal rotation may be applied and interpreted instead. Although the researcher should always report

justification for the rotation method chosen, based on theoretical or empirical evidence, only one-quarter of the analyses in this review provided such justification. Further, some analyses employed orthogonal rotations despite evidence to suggest correlations between factors. Loehlin (1998) indicated use of an orthogonal rotation with correlated factors leads to inflated factor loadings that may influence the interpreted solution. Previous reviews of factor analysis consistently found researchers employed adequate to large sample sizes for application in factor analysis studies (Fabrigar et al., 1999; Henson et al., 2004; Henson & Roberts, 2006). However, this review indicates most studies involved sample sizes under 300 participants, which fail to meet recommendations by Tabachnick and Fidell (2007) for a minimum of 300 cases and Comrey and Lee (1992) who suggest samples sizes below 300 are considered fair to poor. Larger sample sizes generally produce more stable factor structures and better approximate population parameters. In addition to absolute sample sizes, participant to item ratios ranging from 3:1 to 10:1 are referenced in the literature as standards (Cattell, 1978; Costello & Osborne, 2005; Everitt, 1975; Gorsuch, 1983; Tinsley & Tinsley, 1987). Therefore, although absolute sample size recommendations were not met, most analyses in this review met the 10:1 recommended participant to item ratio. Other research does suggest "rules of thumb" for sample size are not appropriate because as the quality of the data, including factor saturation (i.e., number of items loading on each factor) and item communalities (i.e., the total amount of variance for an item explained by the extracted factors), improves, large sample sizes become less critical. Therefore, it is generally recommended that authors seek the largest sample size feasible and then examine factor saturation and item communalities to determine whether further data collection is warranted. Evidence of

this process of examining factor saturation and item communalities in view of sample size was not found in this review of medical education instrument development practice.

A combination of multiple criteria, specifically parallel analysis, minimum average partial, and the scree test, is recommended for determining the number of factors to retain in a solution. However, findings from this review suggest nearly half of these decisions were based on only a single criterion. For roughly an additional 10%, the criterion/criteria used were not reported. Consistent with previous reviews of factor analysis (Fabrigar et al., 1999; Henson et al., 2004; Henson & Roberts, 2006; Pohlmann, 2004), Kaiser's eigenvalue greater than one rule, though largely discredited, and Cattell's scree test were most commonly employed. Each of these methods tends to overestimate the number of factors to retain particularly as the number of variables increase. Only a handful of studies made use of minimum average partial or parallel analysis, though it should be noted these tools are not generally included in most statistical software packages, and, therefore, not readily available to most researchers.

Apart from the five key methodological decision points in factor analysis – model of analysis, sample, extraction and rotation method, and criteria for factor retention – other methodological steps are taken in the analysis and need to be reported for the reader, yet this review suggests limited reporting practices. For instance, to best interpret and potentially replicate a factor solution, all factor loadings for all items must be reported in a factor pattern matrix. However, more than one-third of the reviewed analyses failed to provide this complete data, reporting only select loadings, and one-third reported none of the factor loadings. In addition, to understand which items are interpreted as loading on which factors, the minimum factor loading requirement must be

clearly stated, although what minimum is selected is at the discretion of the researcher. Nearly half of the analyses in this study did not provide this information; without reporting this threshold, the reader cannot understand fully the factor structure. Where minimums were reported, 0.40 was most often selected, a minimum considered as fair to poor (Comrey & Lee, 1992) and adequate (Tabachnick & Fidell, 2007). Similarly, authors failed to report the factor eigenvalues and percent variance explained by each factor and by the total solution in roughly half of the analyses in this study. Although a specific threshold has not been established, the overall percent variance explained by the model suggests the utility of the instrument and should be provided to the reader.

Although this was a review of exploratory factor analysis, each instrument study was examined to determine whether a confirmatory factor analysis was more appropriate based on existing theory or the research question. Almost all studies investigated new or substantially revised instruments, indicating the use of exploratory factor analysis as a best first step. Although several studies did expand on the EFA seeking confirmation of the model through CFA, most did not.

Conclusions

Medical education, across the continuum, is an educational system in which most instrument development, apart from national standardized examinations, is conducted at the institutional level, by individuals with varying levels of expertise, operating with little to no funding (Carline, 2004; Cook et al., 2007; Reed et al., 2005; Shea et al., 2004). Yet, this does not preclude this research from the standards for best practice. Evidence from this review suggests efforts are made to seek reliability and validity evidence expected, given the factor analysis research design; however, the evidence also indicates

a large pool of instruments with only limited reliability and validity evidence based on a narrow few sources, specifically content and internal structure. What appears to be lacking is further evidence to indicate how scores on the instrument relate to other theoretically-related or unrelated variables, how scores on the instrument may predict important expected outcomes, or whether scores on the instrument remain stable or change over time as anticipated by the theoretical understanding of the construct. Investigation of these sources of evidence requires time and more detailed research designs, including longitudinal designs; yet, these sources of evidence are critical to the development of a well-rounded argument for reliability and validity of an instrument. Currently, from these instruments with limited supporting evidence, researchers and educators derive important implications about learners across the continuum of medical education including physicians in practice and curricular programs. Researchers are encouraged to work to build bodies of research around these and other existing measurements reported in the literature. Educators and other readers should be cautious, however, in adopting instruments from the literature without careful consideration of the available supporting evidence. Finally, peer reviewers should be asked to promote instrument development research more consistent with best practice through their review and selection of research for publication.

Further, the evidence available to support the internal structure, specifically the evidence based on dimensionality from a factor analysis, often rests on inappropriate methodology or a lack of reporting of methodology to enable the determination of consistency with best practice. Factor analysis is a complex technique with multiple methodological decision points requiring an informed researcher. This review provides

evidence of the gap between current practice and best practice, highlighting the need for extensive development of additional expertise within the research community including medical education researchers and peer-reviewers. Again, researchers are encouraged to review current recommendations for best practice as outlined here and to be cautious in relying on traditional methods published in the literature. Educators and other readers may not be expected to know the intricacies of such a complex statistical technique; therefore, the peer-review process must help ensure sound methodological techniques are applied in the literature on instrument development across the medical education continuum.

Limitations

The findings and conclusions from this study are tempered by the limitations of this review. Specifically, although a careful review of the literature based on clear inclusion criteria was conducted, there stands the potential that articles were not included in the review that met the criteria. However, with a sample of 62 articles across the continuum of medical education, measuring multiple constructs and published in a variety of peer-reviewed journals, the researcher is confident these findings reflect current practice.

The *Standards of Educational and Psychological Testing* (AERA et al., 1999) provided the framework for the review of reliability and validity evidence for this study, a contemporary perspective of validity as a unitary concept derived from five sources of evidence. Although this contemporary perspective should drive medical education instrument development, it is evident in previous literature and this current review that the traditional validity terminology framed by the three types of validity – content

validity, criterion validity, and construct validity – remains predominant in the medical education literature. Although some efforts have been made to communicate the contemporary perspective from the *Standards of Educational and Psychological Testing* as published in 1999 to medical education research practitioners, exposure of these authors to these concepts may be limited and may influence the scope of techniques for establishing validity evidence that are seen present in this current review.

Further, this review was limited to instrument development articles that specifically employed exploratory factor analysis. EFA is a technique most appropriate in the early developmental stages of a new or revised instrument. Therefore, the scope of findings is likely influenced by this fact as researchers may have been less likely to engage in longitudinal analysis or further data collection that would have allowed for investigation of some sources of validity evidence. Finally, this review does not reflect current practice in confirmatory factor analysis in medical education instrument development. Therefore, only conclusions about exploratory factor analysis in medical education instrument development are appropriately reported in the conclusions to this study.

Recommendations for Instrument Development Practice Employing Exploratory Factor Analysis

1. The first step in developing a new instrument or revising an existing instrument for testing in a new population is clearly defining the measured construct with support from theoretical literature and previous empirical investigations.
2. The process of moving from the defined construct to the measured variables, or items for the instrument, must be documented in detail. It is not sufficient to say

item content was derived from the literature or borrowed from existing instruments. Rather, a blueprint of the construct should be developed that communicates the key content areas. Development of the blueprint may involve focus groups, interviews, or observations of the target population; extensive review of the literature; or collaboration with content experts. The process of item development for each content area should be described, including who wrote the items with their qualifications, techniques employed (e.g., Delphi technique or items taken from other instruments), and any pretesting that may occur.

3. When applying an existing instrument to a new population, the items must be reviewed to ensure the construct is fully represented and that all items are relevant to the new population. Revisions to existing items, deletion of items, or development of new items may be necessary. Engaging a sample from the target population in a review of the items through focus groups, interviews, or surveys can provide feedback on the clarity, relevance, and completeness of the items. All instrumentation, whether new or existing, should be reviewed by experts in the measured construct. The researchers should fully describe for the reader the qualifications of these experts and the process of review they undertake. Pilot test items with a sample from the target population to provide a round of testing to examine variability or patterns of non-response that can inform further revisions before the final administration for data analysis and testing.
4. Seek the largest sample size possible. Set a participant to item ratio goal (e.g., 10:1) and examine factor saturation (i.e., number of items loading on each factor) and item communalities (i.e., the total amount of variance for an item explained

by the extracted factors) after initial factor analysis. If data quality is not adequate, engage in further data collection before proceeding with further analysis. If there are concerns about adequacy of the sample size, run a power analysis.

5. Consider all appropriate measures of reliability and do not rely exclusively on internal consistency. The best argument for reliability is based on multiple reliability statistics ruling out individual threats. Plan to collect data from a small subset of the sample in a follow-up administration of the instrument to enable test-retest calculations. Whether test-retest reliability or stability is most appropriate and the appropriate duration between the two administrations depends on the theoretical understanding of the construct; is it a state that is expected to change, or is it a trait that should remain stable? Use theory to guide the selection of this time-period, recognizing researchers must accommodate feasibility concerns. When collecting data from multiple raters, researchers should calculate the inter-rater or intra-rater reliability statistic, since this requires no additional data collection.
6. Do not rely on default settings in statistical software packages or on tradition from previously published literature using exploratory factor analysis in instrument development. Each analysis is unique and methodological decisions must be made based on the construct, the structure of the instrument and items, and the quality of the data. Principal component analysis and orthogonal (varimax) rotation are default settings in most statistical software packages; yet, these techniques are most often *not* appropriate in social science instrument

development research. It is unclear the extent to which these defaults influence extraction and rotation method selection in this and previous reviews. However, it appears to warrant further consideration through future research or potential dialogue between social science researchers and statistical software developers.

7. Principal component analysis retains error variance in the empirical model; therefore, opportunities for generalizability to other samples and contexts, or for further confirmation testing, are limited. Exploratory factor analysis using a common factor model extraction method produces an error free model and is most appropriate for instrument development research. Researchers should consider the item level of measurement (i.e., nominal, ordinal, interval-ratio) when selecting an extraction method and report this method with justification for the reader.
8. Finding a rotation to be an "interpretable" rotation, or one that is meaningful for the researcher based on *a priori* expectations or theory, does not provide sufficient justification for its selection. Although the goal is to achieve a meaningful, interpretable solution, researchers should select a rotation method based on the theoretical and empirical evidence of the correlations between the underlying factors of an instrument. Within the social sciences, an oblique rotation is more likely than an orthogonal rotation to represent accurately the data as factors may correlate with this rotation. Researchers should first apply an oblique rotation, and then examine the factor pattern matrix and factor correlation matrix. If factors are not correlated, then it would be reasonable to select and interpret the orthogonal rotation of the data. Details of this decision-making

process, including correlations between the factors and the exact oblique or orthogonal rotation applied (e.g., varimax, promax, direct oblimin), must be reported.

9. Researchers must employ and report multiple criteria in determining the number of factors to retain, preferably including the use of minimum average partial or parallel analysis, although currently access to these techniques is limited. Researchers should be cautious in placing full faith in the Kaiser eigenvalue greater than one rule and the Cattell scree test, as each tends to overestimate the number of factors to retain. Bear in mind the recommendation for a minimum of three items per factor to achieve factor stability (Floyd & Widamen, 1995; Tabachnick & Fidell, 2007). Ultimately, the factor model needs to be interpretable and congruent with theoretical foundations of the construct; researchers must articulate this relationship between the empirically derived factor structure and the theoretical structure of the construct to provide the reader with supportive evidence for validity.
10. To create opportunities for other researchers or educators to potentially apply or test an instrument with a new sample, the items need to be reported within the publication. Further, the factor loadings for all items on all factors should be provided in a factor pattern matrix. Without such evidence, interpretation of the solution by the reader is constrained. Though various guidelines are available, the minimum factor loading required for an item to load on a factor is ultimately at the discretion of the researcher; however, the key point is that this minimum must be reported for the reader. Otherwise, the factor patterns cannot be fully

understood or replicated. If particular items fail to meet the minimum factor loading threshold, the researcher should use additional item analyses (e.g., item variability, sub-scale alpha-if-item-deleted) to determine whether to recommend further testing to assess the fit of the item within the solution or to advise the reader to drop the item from the instrument in future applications.

11. The eigenvalues for each retained factor, and the percent variance explained by each factor and the total solution should be reported. In exploratory factor analysis, the percent variance explained is the percent of *shared* variance explained by the solution. In principal component analysis, this percentage represents the percent of *total* variance explained. The reader should keep this in mind when evaluating factor analyses using the two different models, EFA and PCA, as these percentages are not comparable. The researcher should be careful to report this appropriately; evidence indicates researchers employ EFA methods and report on the total variance explained. This can be misleading.
12. Researchers should not rely on validity evidence reported in earlier validation studies of an existing instrument. Use this data to inform the current work; however, fully investigate each source of validity evidence to the extent feasible and practicable for each new application. Further, use of a factor analysis does not exclude the researcher from pursuing evidence based on other sources. Though it does suggest the researcher will investigate and report reliability and validity using certain techniques, efforts should be made to extend the supportive evidence beyond that based on internal structure.

13. Researchers and consumers of research must be tentative in drawing conclusions, as an instrument is not valid or invalid, reliable or unreliable based on a single or few investigations. Reliability and validity are not inherent to the instrument. They are an interaction between the instrument, context in which the measurement occurs, and the sample. As Streiner and Norman (2008) state "the most that we can conclude regarding the results of any one particular study is, 'We have shown the scale to be valid *with this group of people and in this context.*'" (p. 251). Researchers should seek evidence to support reliability and validity when any of these three variables vary. If certain sources of evidence for validity cannot be determined in a study, acknowledge this as a limitation and area for future research. When possible, engage in additional data collection with new or diverse samples to allow for further model testing; develop a longitudinal research agenda that makes the investigation of other sources of validity evidence (e.g., predictive or criterion validity evidence) possible to begin to build a body of knowledge around the measurement of a given construct.

Future Research

Further research in two key areas is required to provide the full context to interpret overall instrument development across the continuum of medical education. As this review focused on exploratory factor analysis, much of what was reviewed were instruments in early developmental stages. As mentioned previously, this may constrain the sources of evidence relevant for investigation by the researcher. Therefore, an equivalent review of instrument development studies employing confirmatory factor analysis would provide a more complete picture of factor analysis in instrument

development and provide a wider scope of potential applications of techniques for establishing validity evidence. Further, a look at instrument development more generally, not restrained to factor analysis studies, would provide an even clearer understanding of the consistency of medical education instrument development with best practices.

Although this review was comprehensive in its abstraction of techniques for establishing validity evidence and comparison of these techniques to best practices, more specification is possible and may provide greater richness to the understanding of validity evidence. For example, in future reviews, rather than only documenting that internal consistency was measured and reported using Cronbach's alpha, a researcher might also document the value of alpha. Similarly, though statistical significance was found in some investigations of differences between theoretically relevant groups, a future review might consider the practical significance of these differences, as measured by effect size.

Finally, two primary questions remain: (a) why does the gap between medical education instrument development researchers' current practice and best practices exist?, and (b) what can be done to address this gap to ensure researchers conduct well-informed instrument development grounded in best practices? Most likely a qualitative investigation into this first question will provide insight into next steps for addressing the second question. Future research may involve interviews with medical education research practitioners to understand their educational background and training in instrument development, what resources they have available and have employed in current practice, and what additional resources they feel may provide the necessary support and professional development to bridge the gap between current and best practices. In addition, similar interviews with journal editors and reviewers may provide

further insight as these persons are the gatekeepers for what reaches the published literature. Lastly, examination of the growing number of master degree programs focused in medical education may provide information on the quality and quantity of research training provided through these specialized programs of study to physicians, basic scientists, and other educators working in medical education. One can anticipate that professional development of medical education researchers, potentially situated within existing regional and national conferences, local experts in instrument development who might advise on individual instrument development projects, and accessible, reader-friendly books on best practices targeted to the research practitioner would likely be beneficial. Currently, books on instrument development best practices do exist both generally and specifically for medical education. However, a version that provides designated time and space to explore the complex methodologies of exploratory factor analysis or a version that considers validity evidence through the lens of the contemporary perspective has not yet been identified.

List of References

List of References

References for the Review of the Literature

Albert, M., & Reeves, S. (2010). Setting some new standards in medical education research. *Medical Education, 44*(7), 638-639.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33-40.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for educational and psychological testing*. (1999). Washington, DC: American Educational Research Association.

Andreatta, P.B., & Gruppen, L.D. (2009). Conceptualising and classifying validity evidence for simulation. *Medical Education, 43*(11), 1028-1035.

Artino, A.R., Durning, S.J., Creel, A.H. (2010). AM last page: Reliability and validity in educational measurement. *Academic Medicine, 85*(9), 1545.

Badia, X., Prieto, L., Linacre, J.M. (2002). Differential item and test functioning (DIF & DTF). *Rasch measurement transactions, 16*(3), 883-892. Retrieved on December 16, 2010 from http://www.rasch.org/rmt/contents.htm.

Bandalos, D.L., & Boehm-Kaufman, M.R. (2009). Four common misconceptions in

exploratory factor analysis. In: Lance, C.E., & Vandenberg, R.J. (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences*. New York: Routledge Taylor & Francis Group.

Beckman, T.J., Ghosh, A.K., Cook, D.A., Erwin, P.J., & Mandrekar, J.N. (2004). How reliable are assessments of clinical teaching? *Journal of General Internal Medicine, 19*(9), 971-977.

Bentler, P.M., & Kano, Y. (1990). On the equivalence of factors and components. *Multivariate Behavioral Research, 25*(1), 67-74.

Boulet, J.R., Champlain, A.F., & McKinley, D.W. (2003). Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical Teacher, 25*(3), 245-249.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*(1), 111-150.

Campbell, D.T. & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Carline, J.D. (2004). Funding medical education research: Opportunities and issues. *Academic Medicine, 79*(10): 918-924.

Cattell, R.B. (1966). The scree test or the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cook, D.A., & Beckman, T.J. (2006). Current concepts in validity and reliability for

psychometric instruments: Theory and application. *The American Journal of Medicine, 119*(2), 166.e7-166.e16.

Comrey, A.L., & Lee, H.B. (1992). *A first course in factor analysis*. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Costello, A.B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation, 10*(7), 1-9.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L.J. (1971). Test validation. In *Educational measurement* (ed. R.L. Thorndike). Washington, DC: American Council on Education.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

DeVellis, R.F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Downing, S.M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38*(9), 1006-1012.

Downing, S.M. & Haladyna, T.M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327-333.

Downing, S.M. (2003). Validity: on the meaningful interpretation of assessment data.

Medical Education, 37(9), 830-837.

Dumenci, L., & Achenbach, T.M. (2008). Effects of estimation methods on making trait-level inferences from ordered categorical items for assessing psychopathology. *Psychological Assessment, 20*(1), 55-62.

Eslaminejad, T., Masood, M., Ngah, N.A. (2010). Assessment of instructors' readiness for implementing e-learning in continuing medical education in Iran. *Medical Teacher, 32*, e407-e412.

Everitt, B.S. (1975). Multivariate analysis: The need for data and other problems. *British Journal of Psychiatry*, *126*, 237-240.

Evidence for Policy and Practice Information and Co-ordinating Centre. (2010). *What is a systematic review?* Retrieved on September 23, 2010, from http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=67.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Floyd, F.J., & Widaman, K.F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7*(3), 286-299.

Gorsuch, R. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research*, *25*(1), 33.

Gorsuch, R.L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Green, S., Higgins, J.P.T., Alderson, P., Clarke, M., Mulrow, C.D., & Oxman, A.D. (2008) Chapter 1: Introduction. In: Higgins, J.P.T. & Green, S (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.1 (updated

September 2008). The Cochrane Collaboration, Retrieved on September 23, 2010, from www.cochrane-handbook.org.

Guadagnoli, E., & Velier, W.R. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265-275.

Henson, R.K., Capraro, R.M., & Capraro, M.M. (2004). Reporting practices and use of exploratory factor analyses in educational research journals: Errors and explanation. *Research in the Schools, 11*(2), 61-72.

Henson, R.K., & Roberts, J.K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416.

Higgins J.P.T. & Deeks, J.J. (2008). Chapter 7: Selecting studies and collecting data. In: Higgins, J.P.T. & Green, S. (eds), *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.1 (updated September 2008). The Cochrane Collaboration. Available from www.cochrane-handbook.org.

Hogarty, K.Y., Hines, C.V., Kromrey, J.D., Ferron, J.M., & Mumford, K.R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality and overdetermination. *Educational and Psychological Measurement, 65*(2), 202-226.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Hutchinson, L., Aitken, P., & Hayes, T. (2002). Are medical postgraduate certification processes valid? A systematic review of the published evidence. *Medical Education, 36*(1), 73-91.

Jha, V., Bekker, H.L., Duffy, S.R.G., & Roberts, T.E. (2007). A systematic review of studies assessing and facilitating attitudes towards professionalism in medicine. *Medical Education, 41*(8), 822-829.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Kieffer, K.M. (1999). An introductory primer on the appropriate use of exploratory and confirmatory factor analysis. *Research in the Schools, 6*(2), 75-92.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Kuder, G.F. & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Lam, T.P., Wong, J.G.W.S., Ip, M.S.M., Lam, K.F., & Pang, S.L. (2010). Psychological well-being of interns in Hong Kong: What causes them stress and what helps them. *Medical Teacher, 32*, e120-e126.

Loehlin, J.C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Lubarsky, S., Charlin, B., Cook, D.A., Chalk, C., & van der Vleuten, C.P.M. (2011). Script concordance testing: a review of published validity evidence. *Medical Education, 45*, 329-338.

MacCallum, R.C., & Tucker, L.R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109*, 502-511.

MacCallum, R.C., Widamen, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor

analysis. *Psychological Methods*, *4*, 84-99.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McDonald, R.P. (1985). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McMillan, J.H. (2008). *Assessment essentials for standards-based education*. Thousand Oaks, CA: Corwin Press.

McMillan, J.H. (2007). *Classroom assessment: Principles and practice for effective standards-based instruction* (4th ed.). Boston: Pearson.

Messick, S. (1975). The standard program: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955-956.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Moore, D.E., Green, J.S., & Gallis, H.A. (2009). Achieving desired results and improved outcomes: Integrating planning and assessment throughout learning activities. *The Journal of Continuing Education in the Health Professions*, *29*(1): 1-15.

Mulaik, S.A. (1990). Blurring the distinctions between component analysis and common factor analysis. *Multivariate Behavioral Research, 25*(1), 53-59.

Mulaik, S.A. (2009). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: CRC Press.

Muthén, L.K. and Muthén, B.O. (2010). *Mplus User's Guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Norris, M., & Lecavalier, L. (2010). Evaluating the use of exploratory factor analysis in

developmental disability psychological research. *Journal of Autism and Developmental Disorders, 40*(1), 8-20.

Park, H.S., Dailey, R., Lemus, D. (2002). The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research, 28*(4), 562-577.

Pohlmann, J.T. (2004). Use and interpretation of factor analysis in *The Journal of Educational Research. The Journal of Educational Research, 98*(1), 14-22.

Ratanawongsa, N., Thomas, P.A., Marinopoulos, S.S., Dorman, T., Wilson, L.M., Ashar, B.H., Magaziner, J.L., Miller, R.G., Prokopowicz, G.P., Qayyum, R., & Bass, E.B. (2008). The reported validity and reliability of methods for evaluating continuing medical education: A systematic review. *Academic Medicine, 83*(3), 274-283.

Raykov, T., & Marcoulides, G.A. (2010). *Introduction to psychometric testing*. New York: Taylor & Francis.

Reed, D.A., Cook, D.A., Beckman, T.J., Levine, R.B., Kern, D.E., & Wright, S.M. (2007). Association between funding and quality of published medical education research. *Journal of the American Medical Association, 298*(9), 1002-1009.

Reed, D.A., Kern, D.E., Levine, R.B., & Wright, S.M. (2005). Costs and funding for published medical education research. *Journal of the American Medical Association, 294*(9), 1052-1057.

Reise, S., Waller, N., & Comrey, A. (2000). Factor analysis and scale revision. *Psychological Assessment, 12*(3), 287-297.

Schonemann, P.H. (1990). Facts, fictions, and common sense about factors and

components. *Multivariate Behavioral Research, 25*(1), 47-51.

Schonrock-Adema, J., Heijne-Penninga, M., van Hell, E.A., & Cohen-Schotanus, J. (2009). Necessary steps in factor analysis: Enhancing validation studies of educational instruments. The PHEEM applied to clerks as an example. *Medical Teacher, 31*(6), e226-e232.

Shaneyfelt, T., Baum, K.D., Bell, D., Feldstein, D., Houston, T.K., Kaatz, S., Whelan, C., & Green, M. (2006). Instruments for evaluating education in evidence-based practice. *Journal of the American Medical Association, 296*(9), 1116-1127.

Shea, J.A., Arnold, L., Mann, K.V. (2004). A RIME perspective on the quality and relevance of current and future medical education research. *Academic Medicine, 79*(10), 931-938.

Snook, S.C., & Gorsuch, R.L. (1989). Component analysis versus common factor analysis: A Monte-Carlo study. *Psychological Bulletin, 106*(10), 148-154.

Steiger, J.H. (1990). Some additional thoughts on components, factors, and factor indeterminacy. *Multivariate Behavioral Research, 25*(1), 41-45.

Streiner, D.L., & Norman, G.R. (2008). *Health measurement scales: A practical guide to their development and use*. New York: Oxford University Press.

Tabachnick, B.G. & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson.

Thurstone, L.L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Tian, J., Atkinson, N.L., Portnoy, B., & Gold, R.S. (2007). A systematic review of evaluation in formal continuing medical education. *Journal of Continuing Education in the Health Professions, 27*(1), 16-27.

Tinsley, H.E.A., Tinsley, D.J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology, 34*, 414-424.

Trochim, W.M.K. (2006). Construct validity. Research Methods Knowledge Base. Retrieved on November 22, 2010, from http://www.socialresearchmethods.net/kb/constval.php.

Velicer, W., & Fava, J. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, *3*(2), 231-251.

Velicer, W., & Jackson, D. (1990a). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25*(1), 1-28.

Velicer, W., & Jackson, D. (1990b). Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research, 25*(1), 97-144.

Velicer, W., Peacock, A.C., Jackson, D.N. (1982). A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research, 17,* 371-388.

Veloski, J.J., Fields, S.K., Boex, J.R., & Blank, L.L. (2005). Measuring professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. *Academic Medicine, 80*(4), 366-370.

Widamen, K.F. (2007). Common factors versus components: principals and principles, errors and misconceptions. In R. Cudeck & R.C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177-204). Mahwah, NJ: Lawrence Erlbaum.

Widaman, K.F. (1993). Common Factor Analysis Versus Principal Component Analysis:

Differential Bias in Representing Model Parameters?. *Multivariate Behavioral Research, 28*(3), 263.

Widamen, K.F. (1990). Bias in pattern loadings represented by common factor analysis and component analysis. *Multivariate Behavioral Research, 25*(1), 89-95.

Worthington, R.L., & Whittaker, T.A. (2006). Scale development research: A content analysis and recommendations for best practice. *The Counseling Psychologist*, 34(6), 806-838.

Zwick, W., & Velicer, W. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432-442.

References for Articles Included in the Pilot Study

Beckman, T.J., & Mandrekar, J.N. (2005). The interpersonal, cognitive and efficiency domains of clinical teaching: Construct validity of a multi-dimensional scale. *Medical Education, 39*(12), 1221-1229.

Buck, D.S., Monteiro, F.M., Kneuper, S., Rochon, D., Clark, D.L., Melillo, A., & Volk, R.J. (2005). Design and validation of the Health Professionals' Attitudes Toward the Homeless Inventory (HPATHI). *BMC Medical Education, 5*(2), 1-8.

Durning, S.J., Pangaro, L.N., Lawrence, L.L., Waechter, D., McManigle, J., & Jackson, J.L. (2005). The feasibility, reliability, and validity of a program director's (supervisor's) evaluation form for medical school graduates. *Academic Medicine, 80*(10), 964-968.

Hoban, J.D., Lawson, S.R., Mazmanian, P.E., Best, A.M., & Seibel, H.R. (2005). The

Self-Directed Learning Readiness Scale: A factor analysis study. *Medical Education, 39*(4), 370-379.

Lee, M., Reuben, D.B., & Ferrell, B.A. (2005). Multidimensional attitudes of medical residents and geriatrics fellows toward older people. *Journal of the American Geriatrics Society, 53*, 489-494.

References for Articles Included in the Systematic Review

Aramesh, K., Mohebbi, M., Jessri, M., & Sanagou, M. (2009). Measuring professionalism in residency training programs in Iran. *Medical Teacher, 31*, e356-e361.

Aukes, L.C., Geertsma, J., Cohen-Schotanus, J., Zwierstra, R.P., & Slaets, J.P.J. (2007). The development of a scale to measure personal reflection in medical practice and education. *Medical Teacher, 29*, 177-182.

Boor, K., Scheele, F., van der Vleuten, C.P.M., Scherpbier, A.J.J.A., Teunissen, P.W., & Sijtsma, K. (2007). Psychometric properties of an instrument to measure the clinical learning environment. *Medical Education, 41*, 92-99.

Campbell, C., Lockyer, J., Laidlaw, T., & MacLeod, H. (2007). Assessment of a matched-pair instrument to examine doctor-patient communication skills in practicing doctors. *Medical Education, 41*, 123-129.

Carruthers, S., Lawton, R., Sandars, J., Howe, A., & Perry, M. (2009). Attitudes to patient safety amongst medical students and tutors: Developing a reliable and valid measure. *Medical Teacher, 31*, e370-376.

Chou, B., Bowen, C.W., & Handa, V.L. (2008). Evaluating the competency of

gynecology residents in the operating room: Validation of a new assessment tool. *American Journal of Obstetrics & Gynecology, 199,* 571.e1-571.e5.

Colletti, J.E., Flottemesch, T.J., O'Connell, T.A., Ankel, F.K., & Asplin, B.R. (2010). Developing a standardized faculty evaluation in an emergency medicine residency. *The Journal of Emergency Medicine, 39*(5), 662-668.

Cruess, R., McIlroy, J.H., Cruess, S., Ginsburg, S., & Steinert, Y. (2006). The professionalism mini-evaluation exercise: A preliminary investigation. *Academic Medicine, 81*(10 Suppl), S74-S78.

Di Lillo, M., Cicchetti, A., Lo Scalzo, A., Taroni, F., & Hojat, M. (2009). The Jefferson scale of Physician Empathy: Preliminary psychometrics and group comparisons in Italian physicians. *Academic Medicine, 84*, 1198-1202.

Dimoliatis, I.D.K., Vasilaki, E., Anastassopoulos, P., Ioannidis, J.P.A., & Roff, S. (2010). Validation of the Greek translation of the Dundee Ready Education Environment Measure (DREEM). *Education for Health, 23*(1), 1-16.

Donnon, T., Woloschuk, W., & Mybre, D. (2009). Issues related to medical students' engagement in integrated rural placements: An exploratory factor analysis. *Canadian Journal of Rural Medicine, 14*(3), 105-110.

El-Zubeir, M., Rizk, D.E.E., & Al-Khalil, R.K. (2006). Are senior UAE medical and nursing students ready for interprofessional learning? Validating the RIPL scale in a Middle Eastern context. *Journal of Interprofessional Care, 20*(6), 619-632.

Eslaminejad, T., Masood, M., & Ngah, N.A. (2010). Assessment of instructors' readiness for implementing e-learning in continuing medical education in Iran. *Medical Teacher, 32*, e407-e412.

Flin, R., Patey, R., Jackson, J., Mearns, K., & Dissanayaka, U. (2009). Year 1 medical undergraduates' knowledge of and attitudes to medical error. *Medical Education, 43*, 1147-1155.

Frye, A.W., Sierpina, V.S., Boisaubin, E.V., & Bulik, R.J. (2006). Measuring what medical students think about complementary and alternative medicine (CAM): A pilot study of the Complementary and Alternative Medicine Survey. *Advances in Health Sciences Education, 11*, 19-32.

Gaspar, M.F., Pinto, A.M., da Conceicao, H.C.F., & da Silva, J.A.P. (2008). A questionnaire for listening to students' voices in the assessment of teaching quality in a classical medical school. *Assessment & Evaluation in Higher Education, 33*(4), 445-453.

Gooneratne, I.K., Munasinghe, S.R., Siriwardena, C., Olupeliyawa, A.M., & Karunathilake, I. (2008). Assessment of psychometric properties of a modified PHEEM questionnaire. *Annals, Academy of Medicine, Singapore, 37*, 993-997.

Haidet, P., O'Malley, K.J., Sharf, B.F., Gladney, A.P., Greisinger, A.J., & Street, R.L. (2008). Characterizing explanatory models of illness in healthcare: Development and validation of the CONNECT instrument. *Patient Education and Counseling, 73*, 232-239.

Harlak, H., Dereboy, C., & Gemalmaz, A. (2008). Validation of a Turkish translation of the communication skills attitude scale with Turkish medical students. *Education for Health, 21*(2), 1-11.

Harvey, B.J., Rothman, A.I., & Frecker, R.C. (2006). A confirmatory factor analysis of

the Oddi Continuing Learning Inventory (OCLI). *Adult Education Quarterly, 56*(3), 188-200.

Helayel, P.E., da Conceicao, D.B., da Conceicao, M.J., Boos, G. L., de Toledo, G.B., & de Oliveira Filho, G.R. (2009). The attitude of anesthesiologists and anesthesiology residents of the CET/SBA regarding upper and lower limb nerve blocks. *Revista Brasileira de Anestesiologia, 59*(3), 332-340.

Hendry, G.D., & Ginns, P. (2009). Readiness for self-directed learning: Validation of a new scale with medical students. *Medical Teacher, 31*, 918-920.

Hojat, M., Veloski, J., Nasca, T.J., Erdmann, J.B., & Gonnella, J.S. (2006). Assessing physicians' orientation toward lifelong learning. *Journal of General Internal Medicine, 21*, 931-936.

Hojat, M., Veloski, J., & Gonnella, J.S. (2009). Measurement and correlates of physicians' lifelong learning. *Academic Medicine, 84*(8), 1066-1074.

Holt, K.D., Miller, R.S., Philibert, I., Heard, J.K., & Nasca, T.J. (2010). Residents' perspectives on the learning environment: From the Accreditation Council for Graduate Medical Education Resident Survey. *Academic Medicine, 85*(3), 512-518.

Kane, G.C., Gotta, J.L., Mangione, S., West, S., & Hojat, M. (2007). Jefferson Scale of Patient's Perceptions of Physician Empathy: Preliminary psychometric data. *Croatian Medical Journal, 48,* 81-86.

Kataoka, H.U., Koide, N., Ochi, K., Hojat, M., & Gonnella, J.S. (2009). Measurement of empathy among Japanese medical students: Psychometrics and score differences by gender and level of medical education. *Academic Medicine, 84*(9), 1192-1197.

Klein, B., McCall, L., Austin, D., & Piterman, L. (2007). A psychometric evaluation of the Learning Styles Questionnaire: 40-item version. *British Journal of Educational Technology, 38*(1), 23-32.

Lam, T.P., Wong, J.G.W.S., Ip, M.S.M., Lam, K.F., & Pang, S.L. (2010). Psychological well-being of interns in Hong Kong: What causes them stress and what helps them. *Medical Teacher, 32*, e120-e126.

Leenstra, J.L., Beckman, T.J., Reed, D.A., Mundell, W.C., Thomas, K.G., Krajicek, B.J., Cha, S.S., Kolars, J.C., & McDonald, F.S. (2007). Validation of a method for assessing resident physicians' quality improvement proposals. *Society of General Internal Medicine, 22*, 1330-1334.

Leung, K., & Wang, W. (2008). Validation of the Tutotest in a hybrid problem-based learning curriculum. *Advances in Health Sciences Education, 13*, 469-477.

Lie, D., Bereknyei, S., Braddock, C.H., Encinas, J., Ahearn, S., & Boker, J.R. (2009). Assessing medical students' skills in working with interpreters during patient encounters: A validation study of the Interpreter Scale. *Academic Medicine, 84*(5), 643-650.

Lin, G.A., Beck, D.C., Stewart, A.L., & Garbutt, J.M. (2007). Resident perceptions of the impact of work hour limitations. *Journal of General Internal Medicine, 22*, 969-975.

Lockyer, J.M., Violato, C., Fidler, H., & Alakija, P. (2009). The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. *Archives of Pathology and Laboratory Medicine, 133*, 1301-1308.

McCormack, W.T., Lazarus, C., Stern, D., & Small, P.A. (2007). Peer nomination: A tool

for identifying medical student exemplars in clinical competence and caring, evaluated at three medical schools. *Academic Medicine, 82*(11), 1032-1039.

McLaughlin, K., Vitale, G., Coderre, S., Violato, C., & Wright, B. (2009). Clerkship evaluation – what are we measuring? *Medical Teacher, 31,* e36-e39.

McManus, I.C., Livingston, G., & Katona, C. (2006). The attractions of medicine: The generic motivations of medical school applicants in relation to demography, personality, and achievement. *BMC Medical Education, 6*(11), 1-15.

Menachery, E.P., Knight, A.M., Kolodner, K., & Wright, S.M. (2006). Physician characteristics associated with proficiency in feedback skills. *Journal of General Internal Medicine, 21*, 440-446.

Mihalynuk, T.V., Coombs, J.B., Rosenfeld, M.E., Scott, C.S., & Knopp, R.H. (2008). Survey correlations: Proficiency and adequacy of nutrition training of medical students. *Journal of the American College of Nutrition, 27*(1), 59-64.

Mitchell, R., Regan-Smith, M., Fisher, M.A., Knox, I., & Lambert, D.R. (2009). A new measure of the cognitive, metacognitive, and experiential aspects of residents' learning. *Academic Medicine, 84*(7), 918-926.

Nagraj, S., Wall, D., & Jones, E. (2007). The development and validation of the mini-surgical theatre educational environment measure. *Medical Teacher, 29,* e192-e197.

Orlander, J.D., Wipf, J.E., & Lew, R.A. (2006). Development of a tool to assess the team leadership skills of medical residents. *Medical Education Online, 11*(27), 1-6.

Park, E.R., Chun, M.B.J., Betancourt, J.R., Green, A.R., & Weissman, J.S. (2009).

Measuring residents' perceived preparedness and skillfulness to deliver cross-cultural care. *Society of General Internal Medicine, 24*(9), 1053-1056.

Pentzek, M., Abholz, H.H., Ostapczuk, M., Altiner, A., Wollny, A., & Fuchs, A. (2009). Dementia knowledge among general practitioners: First results and psychometric properties of a new instrument. *International Psychogeriatrics, 21*(6), 1105-1115.

Reinders, M.E., Blankenstein, A.H., Knol, D.L., de Vet, H.C.W., & van Marwijk, H.W.J. (2009). Validity aspects of the patient feedback questionnaire on consultation skills (PFC), a promising learning instrument in medical education. *Patient Education and Counseling, 76*, 202-206.

Riquelme, A., Herrera, C., Aranis, C., Oporto, J., & Padilla, O. (2009). Psychometric analyses and internal consistency of the PHEEM questionnaire to measure the clinical learning environment in the clerkship of a medical school in Chile. *Medical Teacher, 31*, e221-e225.

Rogers, M.E., Creed, P.A., & Searle, J. (2009). The development and validation of social cognitive career theory instruments to measure choice of medical specialty and practice location. *Journal of Career Assessment, 17*(3), 324-337.

Roh, M.S., Hahm, B.J., Lee, D.H., & Suh, D.H. (2010). Evaluation of empathy among Korean medical students: A cross-sectional study using the Korean version of the Jefferson Scale of Physician Empathy. *Teaching and Learning in Medicine, 22*(3), 167-171.

Sargeant, J., Hill, T., & Breau, L. (2010). Development and testing of a scale to assess interprofessional education (IPE) facilitation skills. *Journal of Continuing Education in the Health Professions, 30*(2), 126-131.

Short, L.M., Alpert, E., Harris, J.M., & Surprenant, Z.J. (2006). A tool for measuring physician readiness to manage intimate partner violence. *American Journal of Preventative Medicine, 30*(2), 173-180.

Singer, Y., & Carmel, S. (2009). Teaching end-of-life care to family medicine residents – what do they learn? *Medical Teacher, 31,* e47-e50.

Sladek, R.M., Phillips, P.A., & Bond, M.J. (2008). Measurement properties of the Inventory of Cognitive Bias in Medicine (ICBM). *BMC Medical Informatics and Decision Making, 8*(20), 1-7.

Sodano, S.M., & Richard, G.V. (2009). Construct validity of the medical specialty preference inventory: A critical analysis. *Journal of Vocational Behavior, 74,* 30-37.

Tian, J., Atkinson, N.L., Portnoy, B., & Lowitt, N.R. (2010). The development of a theory-based instrument to evaluate the effectiveness of continuing medical education. *Academic Medicine, 85*(9), 1518-1525.

Tromp, F., Vernooij-Dassen, M., Kramer, A., Grol, R., & Bottema, B. (2010). Behavioural elements of professionalism: Assessment of a fundamental concept in medical care. *Medical Teacher, 32,* e161-e169.

Tsai, T.C., Lin, C.H., Harasym, P.H., & Violato, C. (2007). Students' perception on medical professionalism: The psychometric perspective. *Medical Teacher, 29,* 128-134.

Vasan, N.S., DeFouw, D.O., & Compton, S. (2009). A survey of student perceptions of team-based learning in anatomy curriculum: Favorable views unrelated to grades. *Anatomical Sciences Education, 2,* 150-155.

Vieira, J.E. (2008). The Postgraduate Hospital Educational Environment Measure (PHEEM) questionnaire identifies quality of instruction as a key factor predicting academic achievement. *Clinics, 63*(6), 741-746.

Wall, D., Clapham, M., Riquelme, A., Vieira, J., & Cartmill, R. (2009). Is PHEEM a multi-dimensional instrument? An international perspective. *Medical Teacher, 31*, e521-e527.

Wetzel, A.P., Mazmanian, P.E., Hojat, M., Kreutzer, K.O., Carrico, R.J., Carr, C., Veloski, J., & Rafiq, A. (2010). Measuring medical students' orientation toward lifelong learning: A psychometric evaluation. *Academic Medicine, 85*(Suppl), S41-S44.

Wittich, C.M., Beckman, T.J., Drefahl, M.M., Mandrekar, J.N., Reed, D.A., Krajicek, B.J., Haddad, R.M., McDonald, F.S., Kolars, J.C., & Thomas, K.G. (2010). Validation of a method to measure resident doctors' reflections on quality improvement. *Medical Education, 44*, 248-255.

Wright, S.M., Levine, R.B., Beasley, B., Haidet, P., Gress, T.W., Caccamese, S., Brady, D., Marwaha, A., & Kern, D.E. (2006). Personal growth and its correlates during residency training. *Medical Education, 40*, 737-745.

Appendices

Appendix A. Data Extraction Form.

Data Extraction Form

Article Title:

_____
_____
_____
_____

Journal:

_____

Volume: _____ Issue: _____ Page Numbers: _____

Authors:

_____
_____
_____

Year: _____

Coder:

_____

Construct measured and instrument title (if applicable):

_____
_____
_____

Research design:

_____
_____
_____

NOTES:

Section I: Educational Outcome Level (using Moore et al. 2009 Outcomes Framework)

Level 1: Participation	
Level 2: Satisfaction	
Level 3A: Learning: Declarative Knowledge	
Level 3B: Learning: Procedural Knowledge	
Level 4: Competence	
Level 5: Performance	
Level 6: Patient Health	
Level 7: Community Health	
Not reported	
Unclear	

Section II: Factor Analysis Methodological Decisions and Reported Evidence

A. Sample

	Factor analysis 1	Factor analysis 2
Reported total *n*		
Ratio of number of participants per variable		
Not reported		
Unclear		
Not applicable		

B. Model of Analysis

	Factor analysis 1	Factor analysis 2
Principal Component Analysis (PCA)		
Exploratory Factor Analysis (EFA)		
Not reported		
Unclear		
Does it appear the model was incorrectly labeled? (If yes, describe.)	Y / N / Unclear / NA	Y / N / Unclear / NA

C. Extraction Method

	Factor analysis 1	Factor analysis 2
Principal Component Analysis		
Maximum Likelihood		
Principal Axis Factoring		
Generalized Least Squares		
Other (Please list.)		
Combination (Please specify each method.)		
Not reported		
Unclear		
Was a justification for extraction method reported based on items' level of measurement?	Y / N / NA	Y / N / NA

D. Rotation Method

	Factor analysis 1	Factor analysis 2
Orthogonal		
Which orthogonal rotation was used?		
Oblique		
Which oblique rotation was used?		
If oblique, what coefficients were reported?	Factor correlation only Factor pattern/loadings only Both Unclear None	Factor correlation only Factor pattern/loadings only Both Unclear None
Both orthogonal and oblique (Please specify rotation methods and circle the rotation that was interpreted.)		
Not reported		
Unclear		
None		
Was a justification for the rotation method reported based on hypothesized or theorized relationships between factors?	Y / N / NA	Y / N / NA
Notes:		

E. Criteria for factor retention

	Factor analysis 1	Factor analysis 2
Previous theory		
Number of factors set *a priori*		
Eigenvalue greater than one rule		
Scree test		
Minimum average partial (MAP)		
Parallel analysis (PA)		
Minimum proportion of variance accounted for by factor		
Number of items per factor		
Conceptual interpretability/meaningfulness		
Not reported		
Unclear		
Other (Please describe.)		

F. Item Retention

	Factor analysis 1	Factor analysis 2
Total number of items in the instrument		
Number of factors retained		
List the number of items for each factor separated by a comma (e.g., 4, 6, 3)		

G. Factor loadings

	Factor analysis 1	Factor analysis 2
Minimum factor loading required for an item to load on a factor		
Not reported		
If no minimum cutoff, please indicate lowest factor loading retained on a factor in the solution.		
Unclear		
Which factor loadings were reported?	All factor loadings for all items Only factor loadings meeting the minimum factor loading criteria and/or only factor loadings for the factor the item is designated as loading on None	All factor loadings for all items Only factor loadings meeting the minimum factor loading criteria and/or only factor loadings for the factor the item is designated as loading on None

H. Other reporting expectations

	Factor analysis 1	Factor analysis 2
Were eigenvalues reported for each retained factor?	Y / N	Y / N
Was the % variance explained reported? By factor By total solution	 Y / N Y / N / NA	 Y / N Y / N / NA

I. Was a confirmatory factor analysis (CFA) warranted?

	Factor analysis 1	Factor analysis 2
Yes, this was not a new measure of a new population.		
Yes, but both EFA and CFA were done in the study.		
No, this was a newly developed measure.		
No, this measure was substantially revised or tested in a new population.		

If CFA was warranted, what reasons were given for not using CFA?	Sample size No strong theory Other Not addressed	Sample size No strong theory Other Not addressed

Section III: Other Techniques for Establishing Validity Evidence

Evidence based on Test Content	Face validity	
	Content validity	
	Expert review	
Evidence based on relationships with other variables	Concurrent criterion validity	
	Predictive criterion validity	
	Convergent evidence	
	Discriminant evidence	
	Divergent evidence	
Evidence based on response process	Intra-rater reliability	
	Inter-rater reliability	
	Test-retest reliability	
	Test-retest stability	
	Alternative-form reliability	
	Questioning test takers about process of response to items (e.g., cognitive interviewing)	
Evidence based on internal structure	Internal consistency	
	Dimensionality (factor analysis)	
	Item analysis	
	Differential Item/Test Functioning	
Evidence based on consequences of testing	Differential Item/Test Functioning	
	Other	
Pilot test (If used, please include techniques that were used specifically in the pilot test within this overall table)	N for the pilot test:	

Appendix B. Data Extraction Form – Coding manual.

Coding Manual

Preliminary Information:

Preliminary information provides a systematic way, as recommended by the Cochrane Collaboration, to capture important data about the article itself to enable detailed description of the sample. In particular, title, journal, authors, year, and other basic information should be documented. In addition, the construct being measured should be described and the title of the instrument should be specified (if applicable); these data will help to understand the scope of knowledge, skills, and attitudes being assessed and evaluated across the continuum of medical education and whether singular instruments are being revised and tested in multiple settings or with different populations. Finally, some studies that meet the eligibility criteria focus exclusively on the development and validation of the instrument. However, some studies may describe the instrument development process that led to the measure used in a different research design (e.g., factor scores used in a regression analysis). If the study is focused on instrument development, just write "instrument development". Otherwise, document the problem statement or research question and proposed data analysis to capture how the instrument is being applied in further research.

Section I: Educational Outcome Level (using Moore et al. 2009 Outcomes Framework)

Data from this section will be used to organize output as a filter to determine whether implementation of best practices varies at different outcome levels. Please place an X in the box to indicate at what educational outcome level the instrument assessed or evaluated. The description, data sources, and methods provided below are to assist in distinguishing between levels. If more than one instrument is used in the article, please complete a data extraction form for each instrument.

Outcomes Framework	Description	Data Sources and Methods
Participation LEVEL 1	Number of learners who participate in the educational activity	Attendance records
Satisfaction LEVEL 2	Degree to which expectations of participants were met regarding the setting and delivery of the educational activity	Questionnaires/surveys completed by attendees after an educational activity
Learning: Declarative Knowledge LEVEL 3A	The degree to which participants state *what* the educational activity intended them to know	Objective: Pre- and post-tests of knowledge Subjective: Self-report of knowledge gain

Learning: Procedural Knowledge LEVEL 3B	The degree to which participants state *how* to do what the educational activity intended them to know how to do	Objective: Pre- and post-tests of knowledge Subjective: Self-report of knowledge gain (e.g., reflective journal)
Competence LEVEL 4	The degree to which participants *show* in an educational setting *how* to do what the educational activity intended them to be able to do	Objective: Observation in educational setting (e.g., online peer assessment and EHR chart simulated recall) Subjective: Self-report of competence; intention to change
Performance LEVEL 5	The degree to which participants *do* what the educational activity intended them to be able to do in their practices	Objective: Observed performance in clinical setting; patient charts; administrative databases Subjective: Self-report of performance
Patient health LEVEL 6	The degree to which the health status of patients improves due to changes in the practice behavior of participants	Objective: Health status measures recorded in patient charts of administrative databases Subjective: Patient self-report of health status
Community health LEVEL 7	The degree to which the health status of a community of patients changes due to changes in the practice behavior of participants	Objective: Epidemiological data and reports Subjective: Community self-report

Source: Moore et al. (2009)

Section II: Factor Analysis Methodological Decisions and Reported Evidence

If a study includes more than one factor analysis on the SAME sample, only the factor analysis methods and results for the FA used to draw conclusions will be mapped onto the data extraction form. However, if the study includes more than one factor analysis based on multiple samples or a divided sample (where participants are not repeated in both analyses), data extraction will occur for both FA's using the dual columns on the form.

A: Sample

An instrument development study may include more than one sample – one for developmental stages, or a pilot study, and one for the factor analysis. For this review, the focus is on sample size just in the factor analysis. If more than one FA is conducted, please list the individual sample sizes separated by a comma.

A researcher may choose to present data on factor analysis sample size in one of two ways. First, they may state the total *n* included in the analysis. If they choose to report both the *n* number of respondents and the *n* number of respondents' data included in the factor analysis, please document the latter, the *n* number of respondents' data included in the factor analysis (for example, in the case of missing data that is deleted listwise). Second, they may indicate the ratio of the number of participants per variable. There are various recommendations for minimum sample sizes and ratios, and research suggests data quality can interact with sample size to influence the factor solution. For the data extraction phase, we are not seeking to evaluate sample size but to capture how and what is reported in the factor analysis studies.

Please fill in the box with the appropriate numeric expression used to communicate the sample size in the article.

B: Model of Analysis

Principal component analysis (PCA) and exploratory factor analysis (EFA) are sometimes used interchangeably; however, they are distinctly different models that serve specific research questions. If the goal is data reduction, PCA is more appropriate. Otherwise, if the researcher seeks to identify latent variables, EFA should be performed. For this section, we want to extract which model was *reportedly* used, if reported. The goal here is to capture what model the authors report using and then to document if it appears the model has been incorrectly labeled, such as in these next two examples. Some researchers may state that they conducted an EFA, but they then describe components or total variance, or other terms denoting PCA. Others may say they conducted an exploratory factor analysis or factor analysis, and then say they used principal component analysis as the method. However, please indicate what model they *reportedly* used. Please only document Principal Component Analysis or Exploratory Factor Analysis if they use this phrasing exactly. Otherwise, this would be defined as "Not Reported". A selection of "Unclear" would be made if the authors appear to use the two phrases, EFA and PCA, interchangeably in describing the methods.

Please place an X in the appropriate box and circle Y or N or Unclear to indicate whether, based on available information, the model of analysis was incorrectly labeled.If no model was reported, select NA for this option. Use the notes box to describe any errors made in the selection of model.

C: Extraction method

Please indicate which extraction method was applied. The extraction method should match with the paradigm for the model of analysis reported previously; however, evaluation of any discrepancies will be made by the lead researcher after data extraction is complete as part of the analysis. If only PCA is mentioned, this should be coded as the model of analysis and extraction method.

Please select Y if the justification for selection of the extraction method reflects consideration of the items' level of measurement. Circle N if there is no justification based on

the level of measurement. Finally, if the extraction method was not reported, select N/A for this option.

D: Rotation method

The two main categories of rotation methods are orthogonal and oblique. There are specific rotation methods within each of these main categories. Orthogonal rotations do not allow factors to correlate; whereas, oblique rotations do allow factors to correlate. Varimax is the most common orthogonal rotation, and oblimin and promax are popular oblique rotations. For oblique rotations, both the factor and structure matrices should be reported.

Please place an X to indicate whether an orthogonal, oblique, or both orthogonal and oblique rotations were applied. If the specific rotation type is named, please write out the specific orthogonal or oblique rotation method of write "not reported". If an oblique rotation was applied, please circle which coefficients were reported – factor correlation only, factor pattern only, both, unclear, or none. Circle Y or N to indicate whether justification for the rotation method was reported. If the rotation method was not reported, select NA for this option.

E: Criteria for factor retention

Multiple criteria exist to support the researcher in determining the number of factors to retain in a model, each with more or less potential for accuracy. Please reference the description of each approach in chapter two if detail on each approach is required to appropriately extract this information.

Please place an X to indicate which criteria were reportedly used to determine the retention of factors. If you select other, please describe the criterion used.

F: Item Retention

Please indicate the total number of items included in the instrument. If a pilot study was conducted, list the number of items included in the revised version used for the validation study. Also, indicate the number of factors retained in the model. Finally, list the number of items retained for each factor, using a comma to separate each factor. For example, if factor 1 has 6 items, factor 2 has 4 items, and factor 3 has 10 items, code this as (6,4,10).

G: Factor loadings

There is no commonly accepted recommendation for the minimum factor loading required for an item to load on a factor; selection of a minimum is at the discretion of the researcher. However, it is an expectation that this value will be reported and that all factor loadings for all items will be reported.

Please document the minimum factor loading required for an item to load on a factor in this study. If this information was not reported, write "not reported" and then document the lowest factor loading interpreted as loading on a factor in the solution. If they report another

means of determining which items load on each factor other than using a minimum value, please document this in detail. Next, please indicate which factor loadings were reported: all factor loadings for all items, only factor loadings meeting the minimum factor loading criteria, none.

H: Other reporting expectations

Please circle Y or N to indicate whether eigenvalues for each retained factor were reported in the article. Also, circle Y or N to document whether the percentage of variance explained by each factor and by the total solution was reported. If the factor analysis identifies a uni-dimensional construct, then document whether the eigenvalue and variance explained for the single factor are reported and select N/A for variance explained by total solution.

I. Was a CFA warranted?

If an instrument has already been developed using EFA in a prior study, a CFA is generally appropriate as the next step in producing further evidence for validity by testing the fit of the factor structure to a new data set. However, if an instrument is new or has been substantially revised or if the instrument is being applied with a new population, an EFA is the appropriate technique. In some instances, the sample size will be large enough that a researcher will choose to conduct both an EFA and CFA by splitting their sample into two smaller, equivalent samples.

Please use an X to denote whether a CFA was warranted in the study in lieu of an EFA using the first four options. If a CFA was appropriate but not performed, there may be reasons why the researcher chose to do an EFA. Please document what, if any, reasons the researcher reported for why a CFA was not used.

Section III: Other Techniques for Establishing Validity Evidence – the traditional classification system mapped to the contemporary definition from the *Standards* (1999)

Please place an X in the appropriate box to indicate which "types" of reliability and validity, as they are understood in the traditional classification system for validity, are reported in the article. The goal is to capture accurately what they are actually doing. However, if an author reports using one technique, but uses terminology incorrectly, code the technique in the correct category, and document in the notes section of the form. If there are multiple errors in using the validity and reliability terminology, this would warrant space in the results and discussion sections. Please describe any techniques used to establish evidence for validity based on consequences of testing. Also, if another technique that is not listed is used, select Other and describe the method. If a pilot test was conducted on the preliminary instrument, please check this box. Any techniques used specifically as part of the pilot study will be captured in the same overall table because for reporting purposes we want to be able to communicate overall what techniques are being applied, and the differentiation between techniques used in the pilot study versus the overall study is not needed as it is all part of the instrument development.

Be sure to note all efforts to seek validity evidence for the instrument, even if the findings are not confirming; we are documenting what techniques were applied, not the quality of the

results. Reference the following table and information in chapter two for definitions and more detailed descriptions of the five sources of validity evidence and the traditional validity terms.

Table 2. Comparison of traditional and contemporary approaches to validity evidence

Traditional classification of validity or reliability	Definition	Mapping of traditional to contemporary approach to validity evidence
Construct validity	Degree to which a measure assesses the theoretical construct intended to be measured	"Validity is a unitary concept….All validity is construct validity in this current framework"
Face/content validity	Degree to which an instrument accurately represents the skill or characteristic that it is designed to measure, according to people's experience and available knowledge.	Content validity remains one of five essential sources of evidence, but face validity is no longer considered
Test criterion validity: Concurrent evidence	Degree to which an instrument produces the same results as another accepted, validated, or even "gold standard" instrument that measures the same construct	Relationships with other variables
Test criterion validity: Predictive evidence	Degree to which a measure accurately predicts something it should theoretically be able to predict	Relationships with other variables
Convergent evidence	Degree of agreement between measurements of the same construct obtained by different methodologies (e.g., objective versus subjective)	Relationships with other variables

Discriminant evidence	Degree to which a measure produces results different from the results of another measure of a theoretically unrelated construct	Relationships with other variables
Divergent evidence	Ability of a measure to yield different mean values between relevant groups	Relationships with other variables
Intra-rater reliability	Degree to which measurements are the same when repeated by the same person	Response process
Inter-rater reliability	Degree to which measurements are the same when obtained by different people	Response process
Test-retest reliability	Degree to which the same test produces the same results when repeated under the same conditions (around a two week interval)	Response process
Test-retest stability	Degree to which the same test produces the same results when repeated under the same conditions (around a six month interval)	Response process
Alternative-form reliability	Degree to which alternate forms of the same measurement instrument produce the same results	Response process
Internal consistency (interitem) reliability	How well items reflecting the same construct yield similar results	Internal structure Consequences: absent in the traditional approach

Source: Adapted from Nunnally & Bernstein (1994), Ratanawongsa et al. (2008) and Trochim (2006)

Appendix C. Original Data Extraction Form.

Data Extraction Form

Article Title: _____

Journal: _____

Authors: _____
_____

Year: _____

Coder: _____

NOTES:

Section I: Educational Outcome Level (using Moore et al. 2009 Outcomes Framework)

Level 1: Participation	
Level 2: Satisfaction	
Level 3A: Learning: Declarative Knowledge	
Level 3B: Learning: Procedural Knowledge	
Level 4: Competence	
Level 5: Performance	
Level 6: Patient Health	
Level 7: Community Health	
Not reported	
Unclear	

Section II: Factor Analysis Methodological Decisions and Reported Evidence

A. Sample

Reported total *n*	
Ratio of number of participants per variable	
Not reported	
Unclear	

B. Model of Analysis

Principal Component Analysis (PCA)	
Exploratory Factor Analysis (EFA)	
Not reported	
Unclear	
Was justification for the model reported?	Y / N

C. Extraction Method

Principal Component Analysis	
Maximum Likelihood	
Principal Axis Factoring	
Generalized Least Squares	
Other	
Combination	
Not reported	
Unclear	
Was justification for the method reported?	Y / N

D. Rotation Method

Orthogonal	
Which orthogonal rotation was used?	
Oblique	
Which oblique rotation was used?	
If oblique, what coefficients were reported?	Factor pattern only Structure pattern only Both Unclear None
Both orthogonal and oblique	
Not reported	
Unclear	
None	

Was justification for the rotation method reported?	Y / N

E. Criteria for factor retention

Previous theory	
Number of factors set *a priori*	
Eigenvalue greater than one rule	
Scree test	
Minimum average partial (MAP)	
Parallel analysis (PA)	
Minimum proportion of variance accounted for by factor	
Number of items per factor	
Conceptual interpretability/meaningfulness	
Not reported	
Unclear	

F. Factor loadings

Minimum factor loading required for an item to load on a factor	
Not reported	
Unclear	
Which factor loadings were reported?	All factor loadings for all items Only factor loadings meeting the minimum factor loading criteria None

G. Other reporting expectations

Were eigenvalues reported each retained factor?	Y / N
Was the % variance explained reported?	
By factor	Y / N
By total solution	Y / N

H. Was a confirmatory factor analysis (CFA) warranted?

Yes, this was not a new measure of a new population.	
Yes, but both EFA and CFA were done in the study.	
No, this was a newly developed or substantially revised measure.	
No, this measure was being tested in a new	

population.	
If CFA was warranted, what reasons were given for not using CFA?	Sample size No strong theory Other Not addressed

Section III: Other Techniques for Establishing Validity Evidence

	Construct validity	
Evidence based on Test Content	Face validity	
	Content validity	
	Expert review	
Evidence based on relationships with other variables	Concurrent criterion validity	
	Predictive criterion validity	
	Convergent evidence	
	Discriminant evidence	
Evidence based on response process	Intra-rater reliability	
	Inter-rater reliability	
	Test-retest reliability	
	Equivalence reliability	
	Questioning test takers about process of response to items	
	Records capturing phases on the development of a response	
Evidence based on internal structure	Internal consistency	
	Dimensionality (factor analysis)	
Evidence based on consequences of testing		
Other		

Appendix D. Original Data Extraction Form – Coding manual.

Data Extraction Information

Section I: Educational Outcome Level (using Moore et al. 2009 Outcomes Framework)

Data from this section will be used to organize output as a filter to determine whether implementation of best practices varies at different outcome levels. Please place an X in the box to indicate at what educational outcome level the instrument assessed or evaluated. If more than one instrument is used in the article, please complete a data extraction form for each instrument.

Outcomes Framework	Description	Data Sources and Methods
Participation LEVEL 1	Number of learners who participate in the educational activity	Attendance records
Satisfaction LEVEL 2	Degree to which expectations of participants were met regarding the setting and delivery of the educational activity	Questionnaires/surveys completed by attendees after an educational activity
Learning: Declarative Knowledge LEVEL 3A	The degree to which participants state *what* the educational activity intended them to know	Objective: Pre- and post-tests of knowledge Subjective: Self-report of knowledge gain
Learning: Procedural Knowledge LEVEL 3B	The degree to which participants state *how* to do what the educational activity intended them to know how to do	Objective: Pre- and post-tests of knowledge Subjective: Self-report of knowledge gain (e.g., reflective journal)
Competence LEVEL 4	The degree to which participants *show* in an educational setting *how* to do what the educational activity intended them to be able to do	Objective: Observation in educational setting (e.g., online peer assessment and EHR chart simulated recall) Subjective: Self-report of competence; intention to change

Performance LEVEL 5	The degree to which participants *do* what the educational activity intended them to be able to do in their practices	Objective: Observed performance in clinical setting; patient charts; administrative databases Subjective: Self-report of performance
Patient health LEVEL 6	The degree to which the health status of patients improves due to changes in the practice behavior of participants	Objective: Health status measures recorded in patient charts of administrative databases Subjective: Patient self-report of health status
Community health LEVEL 7	The degree to which the health status of a community of patients changes due to changes in the practice behavior of participants	Objective: Epidemiological data and reports Subjective: Community self-report

Source: Moore et al. (2009)

Section II: Factor Analysis Methodological Decisions and Reported Evidence

A: Sample

A researcher may choose to present data on sample size in one of two ways. First, they may state the total *n* included in the analysis. Second, they may indicate the ratio of the number of participants per variable. There are various recommendations for minimum sample sizes and ratios, and research suggests data quality can interact with sample size to influence the factor solution. For the data extraction phase, we are not seeking to evaluate sample size but to capture how and what is reported in the factor analysis studies.

Please fill in the box with the appropriate numeric expression from the article.

B: Model of Analysis

Principal component analysis (PCA) and exploratory factor analysis (EFA) are sometimes used interchangeably; however, they are distinctly different models that serve specific research questions. If the goal is data reduction, PCA is more appropriate. Otherwise, if the researcher seeks to identify latent variables, EFA should be performed. For this section, we want to extract which model was reportedly used, if reported, and whether justification for how the model fits the research question was provided. It is important to note that researchers may state that they conducted an EFA, but they then describe components or total variance, or other terms denoting PCA. However, please

indicate what model they *reportedly* used. A later evaluation by the lead researcher will seek to capture discrepancies.

Please place an X in the appropriate box and circle Y or N to indicate whether justification was reported.

C: Extraction method

Please indicate which extraction method was applied. The extraction method should match with the paradigm for the model of analysis reported previously; however, evaluation of any discrepancies will be made by the lead researcher after data extraction is complete as part of the analysis.

Please place an X in the appropriate box and circle Y or N to indicate whether justification was reported.

D: Rotation method

The two main categories of rotation methods are orthogonal and oblique. There are specific rotation methods within each of these main categories. Orthogonal rotations do not allow factors to correlate; whereas, oblique rotations do allow factors to correlate. Varimax is the most common orthogonal rotation, and oblimin and promax are popular oblique rotations. For oblique rotations, both the factor and structure matrices should be reported.

Please place an X to indicate whether an orthogonal, oblique, or both orthogonal and oblique rotations were applied. If the specific rotation type is named, please write out the specific orthogonal or oblique rotation method of write "not reported". If an oblique rotation was applied, please circle which coefficients were reported – factor pattern only, structure patter only, both, unclear, or none. Circle Y or N to indicate whether justification for the rotation method was reported.

E: Criteria for factor retention

Multiple criteria exist to support the researcher in determining the number of factors to retain in a model, each with more or less potential for accuracy. Please reference the description of each approach in chapter two if detail on each approach is required to appropriately extract this information.

Please place an X to indicate which criteria were reportedly used to determine the retention of factors.

F: Factor loadings

There is no commonly accepted recommendation for the minimum factor loading required for an item to load on a factor; selection of a minimum is at the discretion of the researcher. However, it is an expectation that this value will be reported and that all factor loadings for all items will be reported.

Please document the minimum factor loading required for an item to load on a factor in this study. If this information was not reported, write "not reported". Next, please indicate which factor loadings were reported: all factor loadings for all items, only factor loadings meeting the minimum factor loading criteria, none.

G: Other reporting expectations

Please circle Y or N to indicate whether eigenvalues for each retained factor were reported in the article. Also, circle Y or N to document whether the percentage of variance explained by each factor and by the total solution was reported.

H. Was a CFA warranted?

If an instrument has already been developed using EFA in a prior study, a CFA is generally appropriate as the next step in producing further evidence for validity by testing model the fit of the factor structure to a new data set. However, if an instrument is new or has been substantially revised or if the instrument is being applied with a new population, an EFA is the appropriate technique.

Please use an X to denote whether a CFA was warranted in the study in lieu of an EFA using the first four options. If a CFA was appropriate but not performed, there may be reasons why the researcher chose to do an EFA. Please document what, if any, reasons the researcher reported for why a CFA was not used.

Section III: Other Techniques for Establishing Validity Evidence – the traditional classification system mapped to the contemporary definition from the *Standards* (1999)

Please place an X in the appropriate box to indicate which "types" of reliability and validity, as they are understood in the traditional classification system for validity, are reported in the article. Please describe any techniques used to establish evidence for validity based on consequences of testing. Also, if another technique that is not listed is used, select Other and describe the method.

Reference the following table and information in chapter two for definitions and more detailed descriptions of the five sources of validity evidence and the traditional validity terms.

Table 2. Comparison of traditional and contemporary approaches to validity evidence

Traditional classification of validity or reliability	Definition	Mapping of traditional to contemporary approach to validity evidence
Face/content validity	Degree to which an instrument accurately represents the skill or characteristic it is designed to measure, according to people's experience and available knowledge	Content validity remains one of five essential sources of evidence, but face validity is no longer considered
Concurrent criterion validity	Degree to which an instrument produces the same results as another accepted or provide instrument that measures the same variable	Relations to other variables
Predictive criterion validity	Degree to which a measure accurately predicts expected outcomes	Relations to other variables
Construct validity	Degree to which a test measures the theoretical construct it intends to measure	"Validity is a unitary concept….All validity is construct validity in this current framework"
Intrarater reliability	Degree to which measurements are the same when repeated by the same person	Response process
Interrater reliability	Degree to which measurements are the same when obtained by different people	Response process
Test-retest reliability	Degree to which the same test produces the same results when repeated under the same conditions	Response process
Equivalence reliability	Degree to which alternate forms of the same measurement instrument produce the same results	Response process
Internal consistency (interitem) reliability	How well items reflecting the same construct yield similar results	Internal structure

	Consequences: absent in the traditional approach

Source: Ratanawongsa et al. (2008)

Appendix E. Data extraction: Coding manual and form development

Section I: January 25, 2011

Preliminary Information

Second coder:
Research design: Not sure exactly what you are looking for here. Experimental / quasi-experimental / non-experimental?

Lead researcher:
This section will help me describe the sample of articles. Specifically, the committee wants to know what types of studies are included – are they solely articles about the development of an instrument? Or do some studies include instrument development and then involve the application of the scores from the instrument to answer further research questions (e.g., regression analysis or a correlation design). See coding manual for extended directions.

Section I

Second coder:
The coding manual distinguishes between types of data sources and methods (e.g., objective vs. subjective) for educational outcome level, but the extraction form only asks for the level(s). Is the source/method important to distinguish or just the level?

Lead researcher:
Differentiation at the level is sufficient. The data sources and methods are provided to serve as examples to help in distinguishing between levels.

Section II

Second coder:

A) Sample: I am assuming you only want the sample size for the study(ies) that utilized factor analysis. This article was a little tricky. I am assuming the sample they used for the FA was the 1029 students, while the 583 were used for validity evidence and the earlier groups were item development/refinement… but this was all a little unclear. It also made me think that there may be articles which include multiple samples in which FA was performed. Maybe need to revise form to include space for multiple samples?

Lead researcher:
Yes, I am interested in the overall sample size used in the factor analysis (1029 students in the Aukes case). However, it is possible they conducted more than one factor analysis

(done sometimes as a 'semi" confirmatory factor analysis). See form and manual for revisions to allow for tracking of multiple samples.

Second coder:

B) Model of Analysis: This article says they used "explorative" factor analysis, which I interpreted as their statement of EFA, and they attempt to provide justification, but I think it's really just a justification for factor analysis, rather than EFA as a choice over PCA. It made me wonder whether the justification category needs to reflect whether the justification is valid or just that they provided one.

Lead researcher:
I had this conversation with Dr. Dumenci. He suggests one can never legitimately justify PCA, as it is never appropriate in instrument development. What we see are people giving justification as to why they are doing an EFA, but it is typically just a way of describing the analysis procedure (e.g., An EFA is appropriate to seek out the underlying dimensions of X instrument). As I think more about this, the key point of this data point is to determine the extent to which PCA is used in place of EFA. Therefore, we need to be able to document what they *reportedly* used and then what their methods indicate they used (in case there are discrepancies). For example, I have read articles where the authors reports in the abstract that exploratory factor analysis was used to identify the underlying dimensions of the construct. However, in the methods section, they go on to say they used principal component analysis with X rotation for the exploratory factor analysis. This is not correct, and this is what we want to capture, if it is occurring. See the form and guide for more.

Second coder:

C) Extraction Method: For the Y/ N justification items, maybe there should be a category for N/A to be used when the method is not reported, or else some instructions to leave blank or circle N if justification is not applicable. (This could also apply to the model of analysis and rotation justification items.)

Lead researcher:
Makes good sense. See form and manual for revision to coding options and directions.

Second coder:

E) Criteria for factor retention: I got a little confused in this article by their use of "substantial criterion", which made me think maybe you'd want to include a category for "Other" after the list of criteria. I also wondered whether to check an item if the authors didn't state it explicitly. In this article, they talked about jumps in explained variance between factors. I wasn't sure whether to interpret this as a "minimum proportion of variance accounted for by factor" (since it wasn't explicitly stated) or to check "unclear" since they seemed to be using this as a criteria, but they didn't give a cut-off value.

Lead researcher:
Yes, I had listed substantial criterion as an "other". I will add this option to the form. I don't think Aukes et al are explicit enough for us to say that they are using minimum proportion of variance accounted for by factor – that would be if they said, "we only retained factors that explained at least 10% of the variance". We could list this as an "other" as well; I think that may be best, so that it is documented. Let's talk about the "unclear" option tomorrow. It was recommended by the Cochrane Collaboration which suggests adding not reported and unclear to all data points. I'm just not able to picture yet when I would use it.

Second coder:

> H) Other reporting expectations: I wondered how to handle this section (eigenvalues for each facto and % variance for each factor) if the article concluded that the items formed a uni-dimensional measure (i.e., only one factor). Maybe an N/A category, along with Y / N for those two items?

Lead researcher:
If a factor analysis is reported, the eigenvalues and percent variance explained should be reported in all instances. If, as in this study, they conclude it is a unidimensional scale, that data is important in supporting the conclusion they made. I did add N/A as an option for reporting variance explained for the total solution; for uni-dimensional scales, it would be redundant b/c the single factor and the total solution are one and the same.

Section III.a

Second coder:
I am not sure I completely understand the difference between the "item analysis" and "differential item/test functioning" categories. Maybe we could go over this tomorrow. I was also wondering how to handle it if the authors use terminology incorrectly. In other words, if they call something one thing and it fits the definition of another, should it be categorized in the way the authors explicitly state it, or should it be marked in the correct category?

Lead researcher:
Yes, let's go over IA and DIF/DTF tomorrow – the latter is a special case of IA, and there is a definition in chapter 2 that might help you. It serves a specific purpose to see if individual items or sets of items or a test perform differently for different populations (e.g males/females, by race). IA might include lots of other things – looking at the item difficulty, item means, s.d., and variances, etc.

For your second point, I think we should be aiming to capture accurately what they are actually doing. So, yes, you would code it as what it actually is; however, this would be an important thing to document in the notes section of the form. If there are multiple errors in using the validity and reliability terminology, this would warrant space in the results and discussion sections.

Section III.b

Second coder:
I had trouble with documenting the pilot study too. I wasn't sure whether the original sample used for item development/reduction in step 2 (350 students / 38 teachers) was considered a pilot, so I did not mark it as such in III.a, nor did I complete III.b.

Lead researcher:
The pilot study table was added based on my pilot study of the 5 articles; however, those revisions came after I had coded all 5 articles, so this was my first attempt to apply it for coding a new article. I think it just complicates things. The point is to understand what techniques are used to establish validity in instrument development. If we have a table for the pilot study and the regular study, then I'll have to report results that way, and I don't really need to report that level of detail. Instead, we will note whether a pilot study was used, but all techniques will be collapsed in one table. See the manual for more definitions.

Section II: First Session – January 26, 2011

1. Reviewed emailed documents dated 1.25.11; there were no questions.
2. The second coder and the lead researcher went through each coding option for the Aukes et al. (2010) article to document agreements and disagreements based on the 1.25.11 version. Disagreements were resolved by consensus. Necessary revisions to the form and coding manual were made:
    a. A notes section was added to part D – Rotation Method to allow for documentation of any errors in the labeling or use of rotations.
    b. The phrase "(e.g., cognitive interviewing)" was added to the validity technique - "questioning test takers about process of response to items"- to improve clarity between this technique and discussion of items with experts or general content validity based on focus groups with target population.

   c. Understanding was confirmed that attitudes can be mapped onto 3A and 3B depending on whether participant states or describes the attitude as was intended by the educational activity.
   d. Further clarification was added for the terms construct and content validity.
3. Using the manual and form (version dated 1.26.11), after updating together during session to reflect above changes, we coded Tian et al (2010) article. We again reviewed our coding to look for agreements and disagreements. There were minimal disagreements; they were resolved through consensus. Again, revisions to the form or manual were made:
   a. If a study includes more than one factor analysis on the SAME sample, only the factor analysis methods and results for the FA used to draw conclusions will be mapped onto the data extraction form. However, if the study includes more than one factor analysis based on multiple samples or a divided sample (where participants are not repeated in both analyses), data extraction will occur for both FA's using the dual columns on the form. The two columns were new to this version of the form.
   b. Under rotation method, if a study uses more than one rotation method, select Combination. A notation was added to prompt the coder to then list the two or more rotation methods used.
   c. For factor loadings, section G, a box was added to capture the lowest factor loading reportedly retained in the solution IF a minimum factor loading required was not provided.
4. These revisions were made to the form, resulting in version 1.26.11b, after the session with Kelly. The updated form and manual were sent to the second coder electronically for use in the next phase of coding three articles - Wright et al., 2006; Frye et al., 2006; Sargeant, 2010 – to be discussed Wednesday, 2/2/11.

Section III: Independent Coding

January 27, 2011

Based on the lead researcher's independent review of the three articles, these minor revisions were made, and then the coding manual and form, dated 1.27.11, were forwarded to Kelly:

1. For Rotation Method: If both orthogonal and oblique is selected, the notation to be sure to document each rotation method type was added.

2. Under Factor Loadings: The phrase “only factor loadings for the factor the item is designated as loading on” was added. This makes the language more consistent with patterns in the studies where they may not have a minimum factor loading cutoff.
3. Divergent validity was added to the framework under relationships with over variables, as suggested in the *Standards* (1999) that considers categorical variables, such as group membership variables where differences in scores on the instrument are anticipated based on theory, to be relevant within this source of validity evidence.

January 31, 2011

Second coder:
In the Wright, et al. (2006) article, I had trouble deciding how to code the rotation method. Clearly, they used both orthogonal (varimax) and oblique (promax), but the results they reported were all related to the varimax rotation, which they justified by interpretability (better separation of factors). I just want to make sure that in cases like this, the intention is to code the method as "Both orthogonal and oblique" and to list the two methods even if they only report results on one of them.

Lead researcher:
This is correct. We should code the method as “Both orthogonal and oblique” and list the two methods. For the Wright et al. (2006) study, I wrote both rotations and circled Varimax to denote it was the method interpreted – this way, I have all of the data around rotations used and interpreted, just in case this becomes important later. I will make a note on the form for this.

Second coder:
In the Frye, et al. (2006) article, the authors never clearly stated how many items were on their final instrument, but I used the information they provided to infer the number of items retained. This was slightly problematic because they appear to have items that overlap on more than one factor. I'm not sure if this is an issue that needs to be addressed on the data extraction form.

Lead researcher:
For this one, I just left the box blank and noted the number of items was “Not Reported”.

February 2, 2011

1. Reviewed three articles (Wright et al., 2006; Frye et al., 2006; Sargeant, 2010), looking at agreements and disagreements for each coding option. Disagreements were resolved through discussion and consensus. Final revisions were made to the form and guide:
    a. An additional phrase was added to Table 2 to clarify the definition for convergent evidence to distinguish it from concurrent criterion evidence.
    b. The format for documenting justification for extraction method was revised to reduce redundancy in data collection.
    c. In the manual, it was clarified that if authors report both the total *n* and the *n* used in the factor analysis (in the case of missing data deleted listwise), we should document the sample size used on the FA.
2. The second coder was provided a hard copy of the six articles to be double-coded for final agreement calculation.
    a. The second coder will scan and return her coded forms to me electronically as she completes them. It was agreed coding should occur sooner rather than later to ensure consistency in application and to keep understandings of the manual "fresh".
3. Following the session, the lead researcher calculated agreement for the three final preliminary articles using the proportion of agreement was agreements divided by agreements plus disagreements. Overall agreement for these three articles was 89.73%.

March 14, 2011

Construct validity was removed from the framework. In trying to interpret the results and make sense of what specific techniques were applied to aid in the development of an argument for validity, the single term "construct validity" lost any meaning as a precise, definable technique. It is recognized that many articles still used this terminology – construct or content validity – however, simply documenting the use of the word left me unable to make sense of precisely what was being done in the study. Definitions for other techniques were specific enough and thorough enough, that I believe all techniques for seeking validity evidence were documented.

Vita

Angela Payne Wetzel was born on June 5, 1981 in Danville, Virginia. She graduated from Tunstall High School, Dry Fork, Virginia, in 1999. She received a Bachelor of Arts degree in Psychology from the University of Virginia in May 2003 and a Master of Education in Adult Education and Human Resource Development from Virginia Commonwealth University in May 2005. Angie served as a graduate assistant for the School of Education teacher certification program in English as a second language (ESL) one year of her master studies and for the School of Education Metropolitan Educational Research Consortium for one year during her doctoral program. She currently has six years of professional experience in medical education, formerly working as the Director of the Curriculum Office for the Virginia Commonwealth University School of Medicine.

Angie presently is a graduate assistant with the Office of Assessment and Evaluation Studies at Virginia Commonwealth University School of Medicine. Publications include collaboration on two book chapters, Locating and Reviewing Related Literature, in *Educational Research: Fundamentals for the Consumer* (6th ed.) published by Pearson Education (in press), and Evaluating Outcomes in Conintuing Education and Training: Theory and Practice, in *International Best Practices for Evaluation in the Health Professions* published by Radcliffe Publishing (in press); first author on a peer-reviewed article titled "Patient safety attitudes and behaviors of

graduating medical students" published in *Evaluation in the Health Professions* (in press); first author of a book review titled "Internet, mail, and mixed-mode surveys: The tailored design method" published in the *Journal of Continuing Education in the Health Professions* (2010); first author a peer-reviewed article titled "Measuring medical students' orientation toward lifelong learning: a psychometric evaluation" published in *Academic Medicine* (2010); first author on an article titled "Establishing a student tutoring program" published in the *Journal of the International Association of Medical Science Educators* (2008). She has presented at several conferences including the *American Association of Medical Colleges* national conference (2011, 2007), the *American Association of Medical Colleges Southern Group on Educational Affairs* regional conference (2011, 2008, 2007), *American Dental Education Association* national conference (2011), *National Evaluation Institute* (2010), and *American Educational Research Association* conference (2010).