Received: 19 August 2022 | Revised: 13 November 2022 | Accepted: 14 November 2022

DOI: 10.1002/aet2.10835

EDUCATOR'S BLUEPRINT

Educator's blueprint: A how-to guide for collecting validity evidence in survey- based research

Jeffery Hill MD, MEd[1] | Kathleen Ogle MD[2] | Michael Gottlieb MD[3] | Sally A. Santen MD, PhD[1,4] | Anthony R. Artino Jr PhD[5]

[1]Department of Emergency Medicine, University of Cincinnati, Cincinnati, Ohio, USA

[2]Department of Emergency Medicine, George Washington University School of Medicine and Health Sciences, Washington, District of Columbia, USA

[3]Department of Emergency Medicine, Rush University Medical Center, Chicago, Illinois, USA

[4]Department of Emergency Medicine, Virginia Commonwealth University School of Medicine, Richmond, Virginia, USA

[5]Department of Health, Human Function, and Rehabilitation Sciences, George Washington University School of Medicine and Health Sciences, Washington, District of Columbia, USA

Correspondence
Jeffery Hill, University of Cincinnati, Cincinnati, OH 45267-0769, USA.
Email: jeffery.hill@uc.edu

Abstract

Surveys are descriptive assessment tools. Like other assessment tools, the validity and reliability of the data obtained from surveys depend, in large part, on the rigor of the development process. Without validity evidence, data from surveys may lack meaning, leading to uncertainty as to how well the survey truly measures the intended constructs. In documenting the evidence for the validity of survey results and their intended use, it is incumbent on the survey creator to have a firm understanding of validity frameworks. Having an understanding of validity evidence and how each step in the survey development process can support the validity argument makes it easier for the researcher to develop, implement, and publish a high-quality survey.

BACKGROUND

Surveys are a common tool used to evaluate educational initiatives and conduct data collection for research. Therefore, it is essential that clinician educators incorporate thoughtful design to ensure meaningful results. The first paper in this series on survey development and implementation focused on the initial steps of survey development: deciding whether a survey is the appropriate tool for assessing variables of interest, settling on the objectives of the survey, picking constructs to assess, and writing and formatting survey items.[1]

However, those steps are insufficient in isolation, and it is equally important to collect evidence to support the validity argument for the survey's results and how those results will be used. There are many steps that can be taken during the development process to establish and properly document validity evidence within a chosen framework. In this paper, we discuss validity frameworks and steps that can be taken in the development process to provide validity evidence.

VALIDITY FRAMEWORKS

In simple terms, validity typically refers to how well the data produced by an assessment tool reflect the intended construct. In other words, does the survey accurately measure what it intends to measure? It is notable that the instruments (in this case, the surveys)

Supervising Editor: Dr. Wendy Coates.

themselves are not said to be "valid," but instead it is the inferences made based on survey results that are more or less valid for a specific purpose, in a specific context, for a specific population, and, often, at a specific time in history.[2] It is also notable that there is no single piece of evidence that can certify a survey's results as "valid." Instead, designers collect *validity evidence* for the survey results and their intended use. In this way, collecting validity evidence is akin to establishing a clinical diagnosis. There is no one clear test that cinches the diagnosis. Instead, through an accumulation of evidence, a case is built to establish the likelihood of a disease.

Over the years, a number of frameworks have been developed to provide structure for assembling a validity argument. These are primarily directed at the validity evidence of assessments and can be applied to survey instruments. The classical validity framework outlines three types of validity (content, criterion, and construct). However, modern interpretations of validity contend that there is only one type of validity—construct validity—and they therefore organize their frameworks under this construct validity banner.[3–5] These modern validity frameworks seek to gather evidence showing a connection between assessments and specific constructs, arguing that assessment results are only useful if they are theoretically and empirically linked to a construct.[5]

Messick's unified validity framework is the most commonly used approach in the medical education literature and has been adopted and modified by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education as part of their *Standards for Educational and Psychological Testing*.[6,7] The Standards define five *sources* of validity: content, response processes, internal structure, relations to other variables, and consequences of testing (see Table 1 for definitions and examples of how authors have documented and demonstrated validity using these five sources of validity evidence).

Educators can also use the validity framework developed by Kane,[3] which is a stepwise, argument-based approach to validity. It has the advantage of relying less on psychometric data and being applicable to both quantitative and qualitative assessments. It is less commonly used in survey research but depending on the role of a survey in an assessment process, it may provide a flexible approach to documenting validity evidence. In this framework, the validity of an assessment is examined methodically starting with a single response item through the final real-world implementation. By collecting and adjudicating validity evidence at each step (scoring, generalization, extrapolation, and implication), assessment developers cohesively build an argument for or against the validity of an assessment's scores and intended use.[8]

COLLECTING EVIDENCE FOR VALIDITY USING MESSICK'S UNIFIED VALIDITY FRAMEWORK

In the development and implementation of a survey, it is important that the survey designer uses methods that allow for the accumulation of validity evidence. Given its widespread use in medical education and its adoption by the measurement community as articulated in the *Standards for Educational and Psychological Testing*, this section will situate the utility of these methods in the context of Messick's unified validity framework.[7] The researcher, however, may use the data from these methods to report validity evidence for their survey results using the framework best suited to their assessment. Deciding on the most appropriate framework is a matter of balancing several factors: (1) which framework is most familiar in a given context, (2) the researcher's own skill and comfort, (3) the type of evidence that can be collected, (4) the complexity of the validity argument, and (5) the stakes of the inferences being made.[3]

We have previously outlined several processes in the early stages of survey development that help build a case for content validity.[1] First, a detailed literature search can help identify previously derived survey instruments assessing the same or similar construct(s) and will help the researcher better refine the phenomena they intend to measure.[9] However, just because a survey has been published does not mean validity evidence has been collected in a rigorous manner.[10] Taking stock of any validity evidence provided in previous publications is important. Researchers should not focus solely on the survey instrument in isolation but instead should consider its intended use in a potentially new context or new population. When applied in a new way, a previously published survey may or may not yield high-quality results that can be used to make valid inferences. For example, a survey that works well for assessing burnout in medical students would need to be tested (and additional validity evidence collected) if that same tool were to be used in a new context (e.g., within a large health care system) or with a new population (e.g., residents). With that said, the efforts at identifying, using, or adapting previously derived surveys add strength to an argument for content validity. Where previously derived instruments are not available, focus groups, expert panels, or Delphi consensus approaches can be used to identify constructs of interest.[11]

After the clear identification of the objectives of the survey and constructs to be assessed, the next step in survey development is to write initial drafts of survey items.[11] Following best practices for item writing, including question wording and response option formatting, helps increase the likelihood that a survey respondent will comprehend and respond to items as intended.[12] The survey items should then be reviewed by content area experts to further collect content validity evidence. In this step, a group of content area experts is tasked with evaluating how well the survey items reflect the constructs, the clarity of the written items, the relevance of the items to the construct(s) being assessed, and the likely distribution of responses.[9] Integrating the feedback from the expert review allows for additional iterative changes to the survey, readying it for the next steps of development: cognitive interviewing and pilot testing.

Cognitive interviewing is a process that allows the survey designer to gather qualitative data on how a small group of respondents actually engage with the prospective survey.[13] This step adds important evidence about response processes as it provides a means to empirically study the way in which survey respondents mentally

TABLE 1 Messick's unified validity framework as articulated in the standards for educational and psychological testing

Category	Definition	Evidence for validity in the context of a survey	Example of validity evidence documentation
Content	The appropriateness of survey content in light of the construct the tool is intended to measure	• Based on previously developed instruments • Robust derivation of survey items stemming from well-defined constructs of interest • Focus groups • Expert reviews of draft items	Padela et al.[19] describe a comprehensive literature review, searching multiple databases. They clearly describe the expertise of the consensus panel that arrived at the constructs of interest. An expert panel reviewed the draft survey "to assure that the clinical vignettes were specific and realistic, to reduce ambiguity within question stems."
Response processes	The psychological processes or cognitive operations of survey takers and the "detailed nature of the performance … actually engaged in" while completing the survey[6]	• Following best practices for question item formulation • Cognitive interviewing of respondents	Example of cognitive interviewing: "We performed cognitive, 'think-aloud' interviews with two senior medical students who had recently matched in EM and two rising fourth-year medical students pursuing non-EM."[20]
Internal structure	The relationships among survey items or sections of a survey, including score consistency/reliability and subscale structure	• Reliability calculations (Cronbach's alpha, inter-rater reliability, factor analysis, generalizability theory) • Data can come from pilot testing but should also be calculated after full data collection	Pickett et al. report Cronbach's alpha for their survey of EM resident training in psychobehavioral conditions[15]
Relationship to other variables	The associations (positive or negative) between the survey scores and data on other variables	• Associations between survey scores and external variables with theoretical associations • Data from pilot testing	"As expected, confidence levels generally increased by PGY; however, this was not always the case among PGY-4s whose confidence ratings were lower than the mean across several skill areas."[1]
Consequences of testing	The positive or negative, intended or unintended effects of survey use	• Behavioral changes as a result of survey administration (mere-measurements effects)[21] • Iterative curricular improvements with post-course surveys • Psychological impact of surveying sensitive constructs	Rarely included as a component of initial published works.[21] May require follow-up studies to assess.

Note: This table provides definitions for the five sources of validity in Messick's framework. Examples of how a survey designer can collect evidence for validity in a given domain are also provided.

process individual survey items, respond using the given response categories, and work their way through the overall survey instrument.[8] As a qualitative technique, cognitive interviews rely on in-depth interviews with a relatively small sample of volunteers whose characteristics resemble those of the ultimate survey sample.

Two common methods for conducting cognitive interviews are the think-aloud and verbal probing techniques. In the think-aloud technique, the interviewer encourages the respondent to verbalize all of their thoughts as they answer each individual survey item. The transcripts from these interviews can then be analyzed to assess whether or not the respondents are engaging with the survey items as intended. In the verbal probing technique, the interviewer asks directed questions of the respondent. These questions can vary in timing and form. They can occur concurrently to the participant responding to survey items, in a retrospective fashion after the respondent has completed a section of the survey, or during natural breaks in the process of filling out the survey (immediate retrospective probing).[9] The format of the questions in this process may include asking the respondent to paraphrase/restate a survey item, describe what they understand of the intent of an item, and defend why they answered in a particular way (among other questions).[13] Taken together, cognitive interviews are a valuable way to both verify problems that survey designers suspect might cause difficulties (e.g., problems understanding the meanings of certain words) and discover unexpected response process challenges (e.g., faulty interpretations that the survey designer did not foresee).

Prior to the full distribution of the survey, it is wise to pilot test the survey on a small group of potential respondents. There is no consensus on the size of the sample used in a pilot test. Instead, the

sample size is usually based on the purpose of the study and the analyses that will be conducted using the pilot data. For example, if reliability or factor analyses are to be conducted, then a larger pilot sample may be needed (e.g., 10 respondents per survey item for factor analysis). Generally, the pilot sample should mirror the intended sample but include individuals outside the group ultimately to be surveyed. Despite one's best efforts, a survey designer never knows exactly how well their instrument will function until it is distributed to potential respondents and completed under typical conditions. This step also allows the researcher to trial the distribution methods and data collection mechanisms, ensuring they function as intended. Finally, pilot testing allows the researcher to test out their analytical approach and begin getting a sense for things like item distribution and variability, internal structure, and sometimes even relations to other variables (on a small scale).

Following pilot testing, or sometime after full distribution of the survey, the researcher should plan to collect data about the internal structure of the survey. For surveys that use scales (i.e., sets of items designed to measure a construct or constructs), factor analysis is often used to investigate the unidimensionality of a given construct, that is, the degree to which the items measure only a single construct. Although the details of factor analysis are beyond the scope of this paper, several useful resources exist.[14] Following factor analysis, the researcher should consider calculating the reliability of the survey scores, using internal consistency reliability calculations between items (i.e., Cronbach's alpha) or test–retest methods. Both factor analysis and reliability analysis help establish validity evidence based on internal structure.

Additionally, the researcher can examine the correlations between survey responses and other variables thought to be related. For example, a survey assessing residents as teachers might expect to find (based on theory and/or prior research) a positive correlation between confidence in bedside teaching skills and advanced levels of training (where senior residents will have had more training and practical experience serving in a supervisory role). By establishing that expected relationships do indeed exist, the researcher adds to the validity argument by showing evidence based on relationship to other variables.

The final form of validity evidence that a survey creator should consider is consequential validity. Consequential validity documents the effects of the survey, including effects of survey administration and the resulting inferences that might be made using those results. Much as we, as clinicians, worry about the downstream consequences of testing (e.g., will this cardiac CT lead to potentially unnecessary invasive testing?), we should also attend to the downstream consequences of assessments, including surveys, in medical education. Indeed, surveys and other assessments should be thought of as interventions with the potential to have intended and/or unintended, beneficial and/or negative effects on those being surveyed, and on educators, researchers, and the larger systems within which the surveys function.[15] For example, a postcourse survey of a novel curriculum that results in iterative improvements to the content, structure, or administration of that curriculum is an example of an intended, beneficial impact of a survey.

It has been shown that administering surveys that assess behavioral intention and socially desirable behaviors can, in some cases, make those behaviors more likely to occur.[16] This change in behavior that results from survey administration is referred to as the mere-measurement effect.[17] Behavioral change that results from survey administration is, however, not always positive. In fact, high-stakes survey instruments may lead to behavioral changes that threaten the validity of the survey data. For example, Appelbaum and colleagues[18] found significant differences in wellness scores on internal surveys versus Accreditation Council for Graduate Medical Education (ACGME) surveys, postulating in part that the differences could have been the result of coaching bias or social desirability bias in respondents taking the ACGME survey, since those scores are used to make high-stakes accreditation decisions.

CONCLUSIONS

In the process of survey development, there are several steps that can be taken to support the validity argument for a survey and its intended use. Building this argument requires a firm understanding of validity frameworks and of how the development process intersects with these frameworks. Though several validity frameworks exist, because of its common use and acceptance, we have focused on Messick's validity framework here, showing how it can be used to guide the development process and support a given validity argument. Evidence based on content is collected through a detailed approach to survey purpose and construct selection. Response processes are examined through an evidence-based approach to response item writing and formatting as well as cognitive interviewing and pilot testing. Internal structure is supported by factor and reliability analysis. Documenting the ways in which survey scores correlate with other known variables extends the validity argument by examining expected relationships. Finally, documenting the intended or unintended, beneficial or harmful effects of survey administration and use further supports the validity argument.

Having made a rigorous validity argument, collected the necessary data, and conducted the appropriate analyses, researchers are now ready to administer their survey. In our next paper, we will cover the challenges of survey administration, focusing on how to reach the intended audience and how to optimize a survey's response rate.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

ORCID

Jeffery Hill https://orcid.org/0000-0001-8369-3022
Kathleen Ogle https://orcid.org/0000-0002-7658-0895
Michael Gottlieb https://orcid.org/0000-0003-3276-8375
Sally A. Santen https://orcid.org/0000-0002-8327-8002
Anthony R. Artino Jr https://orcid.org/0000-0003-2661-7853

TWITTER

Jeffery Hill @@_drjeffy
Kathleen Ogle @@DrKittyKat
Michael Gottlieb @@MGottliebMD
Anthony R. Artino @@mededdoc

REFERENCES

1. Hill J, Ogle K, Santen S, Gottlieb M, Artino A. Educator's blueprint: a "how to" guide for survey design. *Aem Educ Train*. 2022;6:e10796. doi:10.1002/aet2.10796
2. Downing SM, Yudkowsky R, eds. *Assessment in Health Professions Education*. Routledge; 2009. doi:10.4324/9780203880135
3. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560-575. doi:10.1111/medu.12678
4. Phillips A, Durning S, Artino A. *Survey Methods for Medical and Health Professions Education: A Six-Step Approach*. 1st ed. Elsevier; 2021.
5. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e7-166.e16. doi:10.1016/j.amjmed.2005.10.036
6. Validity MS. *Educational Measurement*. 3rd ed. American Council on Education and Macmillan; 1989.
7. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *The Standards for Educational and Psychological Testing*. American Psychological Association; 2014.
8. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50(1):1-73. doi:10.1111/jedm.12000
9. Artino AR, Rochelle JSL, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE guide No. 87. *Med Teach*. 2014;36(6):463-474. doi:10.3109/0142159x.2014.889814
10. Phillips AW Jr, Artino AR Jr. Lies, damned lies, and surveys. *J Graduate Medical Educ*. 2017;9(6):677-679. doi:10.4300/jgme-d-17-00698.1
11. Gehlbach H, Artino AR Jr, Durning S. AM last page: survey development guidance for medical education researchers. *Acad Med*. 2010;85(5):925. doi:10.1097/acm.0b013e3181dd3e88
12. Gehlbach H, Artino AR. The survey checklist (manifesto). *Acad Med*. 2018;93(3):360-366. doi:10.1097/acm.0000000000002083
13. Willis GB, Artino AR Jr. What do our respondents think we're asking? Using cognitive interviewing to improve medical education surveys. *J Graduate Medical Educ*. 2013;5(3):353-356. doi:10.4300/jgme-d-13-00154.1
14. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. *Psychol Assessment*. 1995;7(3):286-299. doi:10.1037/1040-3590.7.3.286
15. Pickett J, Haas MRC, Fix ML, et al. Training in the management of psychobehavioral conditions: a needs assessment survey of emergency medicine residents. *Aem Educ Train*. 2019;3(4):365-374. doi:10.1002/aet2.10377
16. Godin G, Sheeran P, Conner M, et al. Which survey questions change behavior? Randomized controlled trial of mere measurement interventions. *Health Psychol*. 2010;29(6):636-644. doi:10.1037/a0021131
17. Morwitz VG, Fitzsimons GJ. The mere-measurement effect: why does measuring intentions change actual behavior? *J Consum Psychol*. 2004;14(1–2):64-74. doi:10.1207/s15327663jcp1401&2_8
18. Appelbaum NP, Santen SA, Vota S, Wingfield L, Sabo R, Yaghmour N. Threats to reliability and validity with resident wellness surveying efforts. *J Graduate Medical Educ*. 2019;11(5):543-549. doi:10.4300/jgme-d-19-00216.1
19. Padela AI, Davis J, Hall S, Dorey A, Asher S. Are emergency medicine residents prepared to meet the ethical challenges of clinical practice? Findings from an exploratory National Survey. *Aem Educ Train*. 2018;2(4):301-309. doi:10.1002/aet2.10120
20. Jauregui J, Kessler R, Villalón N, et al. Medical student experiences of applying into emergency medicine during the COVID-19 pandemic: a multi-institutional survey of emergency medicine–bound medical students. *Aem Educ Train*. 2021;5(2):e10587. doi:10.1002/aet2.10587
21. Cook DA, Lineberry M. Consequences validity evidence. *Acad Med*. 2016;91(6):785-795. doi:10.1097/acm.0000000000001114

How to cite this article: Hill J, Ogle K, Gottlieb M, Santen SA, Artino AR Jr. Educator's blueprint: A how-to guide for collecting validity evidence in survey- based research. *AEM Educ Train*. 2022;6:e10835. doi:10.1002/aet2.10835