You Can't Fix by Analysis What You've Spoiled by Design: Developing Survey Instruments and Collecting Validity Evidence

Gretchen Rickards, MD
Charles Magee, MD, MPH
Anthony R. Artino Jr, PhD

Surveys are frequently used in graduate medical education (GME). Examples include resident satisfaction surveys, resident work-hour questionnaires, trainee self-assessments, and end-of-rotation evaluations. Survey instruments are also widely used in GME research. A review of the last 7 issues of *JGME* indicates that of the 64 articles categorized as *Original Research*, 50 (77%) included surveys as part of the study design.

Despite the many uses of surveys in GME, the medical education literature provides limited guidance on survey design,[1] and many surveys fail to use a rigorous methodology or best practices in survey design.[2] As a result, the reliability and validity of many medical education surveys are uncertain. When surveys are not well designed, the data obtained from them may not be reproducible and may fail to capture the essence of the attitude, opinion, or behavior the survey developer is attempting to measure. A plethora of factors affecting reliability and validity in surveys includes, but is not limited to, poor question wording, confusing question layout, and inadequate response options. Ultimately, these problems negatively impact the reliability and validity of survey data, making it difficult to draw useful conclusions.[3,4] With these problems in mind, the aim of the present editorial is to outline a systematic process for developing and collecting reliability and validity evidence for survey instruments used in GME and GME research.

The term *survey* is quite broad and could include questions used in a phone interview, the set of items used in a focus group, and the items on a self-administered patient survey. In this editorial, we limit our discussion to self-administered surveys, which are also sometimes referred to as questionnaires. The goals of any good questionnaire should be to develop a set of items that every respondent will interpret the same way, respond to accurately, and be willing and motivated to answer. The 6 questions below, although not intended to address all aspects of survey design, are meant to help guide the novice survey developer through the survey design process. Addressing each of these questions systematically will optimize the quality of GME surveys and improve the chances of collecting survey data with evidence of reliability and validity. A graphic depiction of the process described below is presented in the FIGURE.

Question 1: Is a Survey an Appropriate Tool to Help Answer My Research Question?

Surveys are good for gathering data about abstract ideas or concepts that are otherwise difficult to quantify, such as opinions, attitudes, and beliefs. Surveys are also useful for collecting information about behaviors that are not directly observable (eg, Internet usage or other off-duty behaviors). Before creating a survey, it is imperative to decide if a survey is actually the best method to address your research question or construct of interest. In the language of survey design, a *construct* is the model, idea, or theory you are attempting to assess. In GME, some of the constructs we are interested in assessing are not directly observable, and so a survey is often a useful research tool. For instance, a survey may be helpful in assessing resident opinions about a procedure curriculum. And while this information may provide insight for curriculum improvement, the objective outcomes of that same curriculum might be best assessed through other means, such as direct observation or examination. Thus, surveys often supplement, rather than replace, other forms of data collection.

The surveys used in GME and GME research often address constructs that are psychological in nature and are not directly observable. Examples of the constructs we often want to measure include things like *motivation*, *satisfaction*, and *perceived learning*. Accordingly, it makes sense to assess these constructs by using a survey scale.

All authors are at Uniformed Services University of the Health Sciences. **Gretchen Rickards, MD,** is Assistant Professor of Medicine; **Charles Magee, MD, MPH,** is Assistant Professor of Medicine; and **Anthony R. Artino Jr, PhD,** is Associate Professor of Medicine and Preventive Medicine & Biometrics.

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of Defense or the U.S. Government.

The title of this paper was adapted from Light RJ, Singer JD, Willett JB. *By Design: Planning Research On Higher Education.* Cambridge, MA: Harvard University Press; 1990.

Corresponding author: Anthony R. Artino Jr, PhD, 4301 Jones Bridge Road, Bethesda, MD 20814, 301.295.3693, anthony.artino@usuhs.edu

DOI: http://dx.doi.org/10.4300/JGME-D-12-00239.1

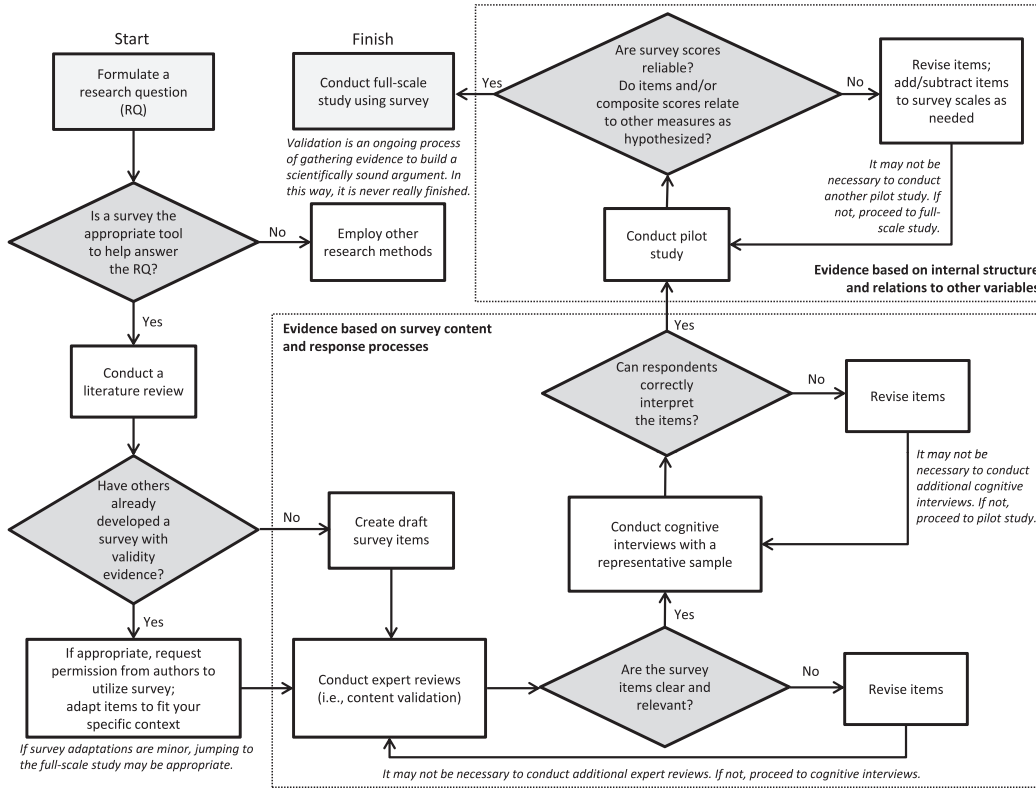

FIGURE **A Systematic Approach to Survey Design for Graduate Medical Education Research**

Survey scales are groups of items on a survey that are designed to assess a particular construct of interest. So, instead of just asking 1 question about *resident satisfaction* (eg, How satisfied were you with the curriculum?), it is often more helpful to ask a series of questions designed to capture the different facets of this satisfaction construct (eg, How satisfied were you with your clinical instructors? How satisfied were you with the teaching facilities? How satisfied were you with the scheduling processes?). Using this approach, an unweighted average score of all the items within a particular scale (ie, a composite score) can be calculated and used in the research study. Generally, the more complex the construct, the more items you will need to create, and thus the longer your survey scale.

Question 2: How Have Others Addressed This Construct in the Past?

A review of the literature can be helpful in this step, both to ensure your construct definition aligns with related research in the field and to identify survey scales or items that could be used or adapted for your purpose.[1] Educators and researchers often prefer to "home grow" their own surveys, yet it may be more useful to review the surveys that already exist in the literature—and that have undergone some level of validation—than to start from scratch. Odds are, if you are interested in measuring a particular construct, someone else has previously attempted to measure it, or something very similar. When this is the case, a request to the authors to adapt their survey for your purposes will usually suffice.

It is important to note, however, that previously validated surveys require the collection of additional reliability and validity evidence in your specific context. Survey validity is the degree to which inferences about the construct measured are appropriate, and validity is sensitive to the survey's target population and the local context. Thus, survey developers collect reliability and validity evidence for their survey in a specified context, with a particular sample, and for a particular purpose. As described in the *Standards for Educational and Psychological Testing*,[5] validity refers to the degree to which evidence and theory support a measure's intended use. The process of validation is the most fundamental consideration in developing and evaluating a measurement tool. This process involves the accumulation of evidence across time, settings, and samples to build a scientifically sound validity argument. Thus, establishing validity is an ongoing process of gathering evidence. In this way, survey validation is

context dependent and, in some sense, is never really finished.[6] Furthermore, it is essential to acknowledge that reliability and validity are not properties of the survey instrument, per se, but of the survey's scores and their interpretations.[5] For example, a survey of student anxiety might be appropriate for assessing aspects of well-being, but such a survey would be inappropriate for selecting the most knowledgeable medical students. In this example, the survey did not change, only the score interpretation changed.

Question 3: How Do I Develop My Survey Items?

Items or questions on a survey should be developed in accordance with the best practices of survey design.[1,7] Writing a well-articulated construct definition is important. As such, interviewing the target population or individuals knowledgeable about the topic can be a useful first step in understanding how others conceptualize or describe your construct of interest. Developing items by using the vocabulary of your target population is also important. For example, instead of asking residents about "the sanitation of slumber facilities," you would likely ask about "the cleanliness of on-call or sleeping rooms." Other key principles of item development include writing questions rather than statements, avoiding negatively worded items, and using response anchors that emphasize the construct being measured rather than using general agreement response anchors.[2,8] Although widely used, general agreement response anchors (eg, strongly disagree, disagree, neutral, agree, strongly agree) are well known to be subject to considerable measurement error.[2]

Once you've drafted your survey items, there are various sources of evidence that might be used to evaluate the validity of your survey and its intended use. These sources have been described in the *Standards for Educational and Psychological Testing*[5] as evidence based on (1) content, (2) response process, (3) internal structure, (4) relationships with other variables, and (5) consequences. Several of the processes described below fit nicely into this taxonomy and are labeled accordingly in the FIGURE.

Question 4: Are the Survey Items Clearly Written and Relevant to the Construct of Interest?

To assess the *content* of your survey, ask experts to review the questions for clarity and relevance to the construct. Experts might include those more experienced in survey design, national content experts, or local colleagues knowledgeable about your topic. The key is to have several experts review the items in detail to ensure the questions "ring true" and adequately cover the construct (or constructs) being assessed. Items that are flagged as ambiguous, cognitively difficult to answer, or minimally related to the construct of interest may require further revisions and repeated expert review.[7] Although beyond the scope of this introductory article, there are several references that outline systematic approaches to conducting an expert review (also known as a content validation).[9,10]

Question 5: Will Respondents Interpret My Items in the Manner That I Intended?

After the experts have assisted in refining the overall survey and specific survey items, it is important to collect evidence of *response process validity* to assess how your study participants interpret your items and response anchors. One means of collecting such evidence is achieved through a process known as cognitive interviewing (or cognitive pretesting).[11] Ideally, cognitive interviewing involves reviewing survey items in detail with a handful of participants who are representative of your target population. This qualitative process generally involves a face-to-face interview during which a respondent reads each question and then explains his or her thought process in selecting an answer. This process is invaluable in identifying problems with question or response wording that may result in misinterpretations or bias. Ultimately, the goal is twofold: to ensure respondents understand the questions as you intended and to verify that different respondents interpret the items the same way and can respond to the items accurately.[7]

Question 6: Are the Scores Obtained From My Survey Items Reliable and Do They Relate to Other Measures as Hypothesized?

The next step is to pilot test your survey and to begin collecting validity evidence based on reliability and relationships with other variables. During pilot testing, members of the target population complete the survey in the planned delivery mode (eg, web-based or paper-based format). The data obtained from the pilot test can then be reviewed to evaluate item range and variance, assess score reliability, and review item and composite score correlations. During this step, descriptive statistics (eg, mean, standard deviation) and histograms that demonstrate the distribution of responses by item are reviewed. This step can provide meaningful evidence of the degree to which individual items are normally distributed and can aid in identifying items that may not be functioning in the way you intended.

Conducting a reliability analysis is another critical step, particularly if your survey consists of several survey scales (ie, several items all designed to assess a given construct, such as *resident satisfaction*). The most common means of assessing scale reliability is by calculating a Cronbach α coefficient. This is a measure of internal consistency reliability; that is, the extent to which the items in your

scale are correlated with one another. Simply speaking, if 5 items on your survey are all designed to measure the construct *resident interest*, for example, then it follows that the 5 items should be moderately to highly correlated with one another. It is worth noting that Cronbach α is also sensitive to scale length. Thus, all other things being equal, a longer survey scale will generally have a higher Cronbach α. Of course, survey length and the concomitant increase in internal consistency reliability must be balanced with the response errors that can occur when surveys become too long and respondents become fatigued. Finally, it is critical to recognize that reliability is a necessary but insufficient condition for validity.[5] That is, to be considered valid, survey scores must first be reliable. However, scores that are reliable are not automatically valid for a given purpose.

Once internal consistency reliability has been assessed, survey developers often create composite scores for each scale. Depending on the research question being addressed, these composite scores can then be used as independent or dependent variables. When attempting to assess unobservable, "fuzzy" constructs—as we often do in GME and GME research—it usually makes more sense to create a composite score for each survey scale than it does to use individual survey items as variables. As described earlier, a composite score is simply an unweighted average of all the items within a particular scale. After composite scores are created for each survey scale, the resulting variables can be examined to determine their relations to other variables you may have collected. The goal in this step is to determine if these associations are consistent with theory and previous research. So, for example, if you created a scale to assess *resident confidence* in a given procedure (eg, lumbar puncture), you might expect the composite variable created from these confidence items to be positively correlated with the volume of lumbar punctures performed (as practice builds confidence) and negatively correlated with procedure-related anxiety (as more confident residents also tend to be less anxious). In this way, you are assessing the validity of the scale items you have created in terms of their relationships to other variables.[5]

Concluding Thoughts

The processes outlined in this editorial are intended to provide a general framework for GME survey development. By following these steps and collecting reliability and validity evidence across time, settings, and samples, GME educators and researchers will improve the quality of surveys as well as the validity of conclusions drawn from surveys. The steps described in this editorial, if they are completed by GME researchers, should be reported in research papers. In future *JGME* editorials we will address several of these processes in greater detail and provide specific guidelines for reporting survey validation findings in *JGME* submissions.

Glossary

Construct—A hypothesized concept, model, idea, or theory (something "constructed") that is believed to exist but cannot be directly observed.

Content validity—Evidence obtained from an analysis of the relationship between a survey instrument's content and the construct it is intended to measure.

Reliability—The extent to which the scores produced by a particular measurement procedure or instrument (eg, a survey) are consistent and reproducible. Reliability is a necessary but insufficient condition for validity.

Response anchors—The named points along a set of answer options (eg, strongly disagree, disagree, neutral, agree, strongly agree).

Response process validity—Evidence obtained from an analysis of how respondents interpret the meaning of a survey and its specific survey items.

Scale—Two or more items intended to measure a construct. Often, however, the word *scale* is used more generally to refer to the entire survey (eg, "a survey scale"). As such, many survey scales are composed of several subscales.

Validity—The degree to which evidence and theory support a measure's intended use.

Validity argument—The process of accumulating evidence to provide a sound scientific basis for the proposed uses of an instrument's scores.

References

1 Gehlbach H, Artino AR, Durning S. AM last page: survey development guidance for medical education researchers. *Acad Med*. 2010;85:925.
2 Dillman D, Smyth J, Christian L. *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. 3rd ed. Hoboken, NJ: Wiley; 2009.
3 Sullivan G. A primer on the validity of assessment instruments. *J Grad Med Educ*. 2011;3(2):119–120.
4 Artino AR, Durning SJ, Creel AH. AM last page: reliability and validity in educational measurement. *Acad Med*. 2010;85:1545.
5 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.
6 Kane MT. *Validation in Educational Measurement*. 4th ed. Westport, CT: American Council on Education/Praeger; 2006.
7 LaRochell J, Hoellein AR, Dyrbe LN, Artino AR. Survey development: what not to avoid. *Acad Intern Med Insight*. 2011;9:10–12.
8 Artino AR, Gehlbach H, Durning SJ. AM last page: avoiding five common pitfalls of survey design. *Acad Med*. 2011;86:1327.
9 Rubio D, Berg-Weger M, Tebb SS, Lee ES, Rauch S. Objectifying content validity: conducting a content validity study in social work research. *Soc Work Res*. 2003;27(2):94–104.
10 McKenzie J, Wood ML, Kotecki JE, Clark JK, Brey RA. Establishing content validity: using qualitative and quantitative steps. *Am J Health Behav*. 1999;23(4):311–318.
11 Willis GB. *Cognitive Interviewing: A tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications; 2005.