You Can't Fix by Analysis What You've Spoiled by Design:
An Introduction to Survey Design for Medical Educators

Workshop Description:

Questionnaires are commonly used in health professions education. Unfortunately, few educators are familiar with the best practices of questionnaire design. The purpose of this workshop is to provide medical educators with an introduction to a systematic process for creating valid and reliable questionnaires that can be used as assessment or research tools.

Upon completion of the workshop, participants will be able to…

1) Recognize how to use a systematic, 7-step process as the framework for questionnaire;
2) Demonstrate how to develop an appropriate set of items (a survey "scale") to characterize the educational construct being measured;
3) Identify common item-writing pitfalls in questionnaire design; and
4) Define the purpose of expert validation, cognitive interviews, and pilot testing.

Presenter:

Dr. Anthony R. Artino, Jr. is an Associate Professor of Preventive Medicine and Biometrics at the Uniformed Services University of the Health Sciences in Bethesda, Maryland. Anthony is also a Commander in the U.S. Navy. He received his Ph.D. in educational psychology from the University of Connecticut and has published 80 peer-reviewed articles on learning and motivation, questionnaire design, and online education. Over the past five years, he has presented more than a dozen questionnaire-design workshops at national and international meetings.

AM Last Page: Survey Development Guidance for Medical Education Researchers

Hunter Gehlbach, PhD, assistant professor of Education, Harvard University; Anthony R. Artino, Jr, PhD, assistant professor of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences; and Steven J. Durning, MD, professor of Medicine, Uniformed Services University of the Health Sciences.

Medical education researchers frequently rely on survey data. For example, of *Academic Medicine*'s 141 research articles from 2009, over half (56%) used surveys. Yet, the literature provides limited guidance on which processes best facilitate the development of surveys—particularly in the design of survey scales (i.e., several items that assess a single underlying construct such as physician empathy or teaching self-efficacy; see example below). This flowchart presents seven steps to facilitate the construction of valid and reliable survey scales.

Step 1

Conduct a ***literature review*** both to ensure that your construct definition aligns with relevant prior research and to identify extant survey scales or items that might be used or adapted for your research context.

Step 2

Conduct ***interviews*** and/or ***focus groups*** to learn how your population of interest conceptualizes and describes your construct of interest.

Step 3

Synthesize the *literature review* and *interview/focus group* data so that the conceptualization of the construct makes theoretical sense to scholars in the field and uses language that your population of interest understands. For example, a scale assessing teaching self-efficacy (i.e., confidence in one's teaching ability) should use words like "confidence in trying out new teaching techniques," not "efficaciousness in experimenting with novel pedagogies."

Step 4

Develop items in accordance with current best practices in survey design.[1,2] For example, the sample scale below uses response anchors that refer to the specific construct (rather than numbers or agree/disagree response anchors).

Step 5

Conduct an ***expert validation*** to assess the items' clarity and relevance to the construct.[3,4]*

Step 6

Conduct ***cognitive pretesting*** through which participants restate each item aloud in their own words as they answer it. This step helps ensure that respondents interpret items in the manner that you intend.[5,6]*

Step 7

Pilot-test your items to check for adequate item variance, reliability, and convergent/discriminant validity with respect to other measures.[7]*

***Note:** After you complete each of these final steps, you may need to revise items and/or repeat steps from this part of the process.

Sample Items From a Teaching Self-Efficacy Scale

1. How confident are you that you can help students remember what they learned in your class?
2. When you need to teach less interesting topics, how confident are you that you can keep all students engaged?
3. How confident are you that you can help students learn when they are unmotivated?
4. How confident are you that you can get through to the most difficult students?

5-point, Likert-type response scale:	Not at all confident	Slightly confident	Moderately confident	Quite confident	Extremely confident

References

1. Dillman DA, Smyth JD, Christian LM. Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method. 3rd ed. Hoboken, NJ: Wiley; 2009.
2. Fowler FJ. Survey Research Methods. 4th ed. Thousand Oaks, Calif: Sage Publications; 2009.
3. McKenzie JF, Wood ML, Kotecki JE, Clark JK, Brey RA. Establishing content validity: Using qualitative and quantitative steps. Am J Health Behav. 1999;23:311–318.
4. Rubio DM, 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;27:94–104.
5. Karabenick SA, Woolley ME, Friedel JM, et al. Cognitive processing of self-report items in educational research: Do they think what we mean? Educ Psychol. 2007;42:139–151.
6. Willis GB. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, Calif: Sage Publications; 2005.
7. DeVellis RF. Scale Development: Theory and Applications. 2nd ed. Newbury Park, Calif: Sage; 2003.

AM Last Page: Avoiding Five Common Pitfalls of Survey Design

Anthony R. Artino, Jr, PhD, assistant professor of preventive medicine and biometrics, Uniformed Services University of the Health Sciences, Hunter Gehlbach, PhD, assistant professor of education, Harvard University, and Steven J. Durning, MD, professor of medicine and pathology, Uniformed Services University of the Health Sciences

Writing good survey items is both an art and a science. Over the last 30 years, scholars have amassed a great deal of scientific evidence on which questionnaire designers can rely.[1–5] The guidelines below present some of the more frequently ignored, but more important, of these survey-design basics.

Pitfall	Survey example(s)	Why it's a problem	Solution(s)	Survey example(s)
Creating a double-barreled item	How often do you talk to your nurses and administrative staff when you have a problem?	Respondents have trouble answering survey items that contain more than one question (and thus could have more than one answer).[1] In this example, respondents who talk to nurses often but talk to administrative staff infrequently will struggle to answer this question. Survey items should address one idea at a time.[1]	When you have multiple questions/premises within a given item, either (1) create multiple items for each question that is important or (2) include only the more important question. Be especially wary of conjunctions in your items.[1,4]	How often do you talk to your nurses when you have a problem? How often do you talk to your administrative staff when you have a problem?
Creating a negatively worded item	In an average week, how many times are you unable to start class on time? The chief resident should not be responsible for denying admission to patients.	Negatively worded survey items are challenging for respondents to comprehend and answer accurately. Double-negatives are particularly problematic and increase measurement error.[1] If a respondent has to say "yes" in order to mean "no" (or "agree" in order to "disagree"), the item is flawed.	Make sure "yes" means yes and "no" means no. This generally means wording items positively.[1]	In an average week, how many times do you start class on time? Should the chief resident be responsible for admitting patients?
Using statements instead of questions	I am confident I can do well in this course. • not at all true • a little bit true • somewhat true • mostly true • completely true	A survey represents a conversation between the surveyor and the respondents. To make sense of survey items, respondents rely on "the tacit assumptions that govern the conduct of conversation in everyday life."[2] Only rarely do people engage in rating statements in their everyday conversations.	Formulate survey items as questions. Questions are more conversational, more straightforward, and easier to process mentally. People are more practiced at responding to them.[1,4]	How confident are you that you can do well in this course? • not at all confident • slightly confident • moderately confident • quite confident • extremely confident
Using agreement response anchors	The high cost of health care is the most important issue in America today. • strongly disagree • disagree • neutral • agree • strongly agree	Agreement response anchors do not emphasize the construct being measured and are prone to acquiescence (i.e., the tendency to endorse any assertion made in an item, regardless of its content).[3] In addition, agreement response anchors may encourage respondents to think through their responses less thoroughly while completing the survey.[4]	Use construct-specific response anchors that emphasize the construct of interest. Doing so reduces acquiescence and keeps respondents focused on the construct in question. Doing so results in less measurement error.[1,4]	How important is the issue of high health care costs in America today? • not at all important • slightly important • moderately important • quite important • extremely important
Using too few or too many response anchors	How useful was your medical school training in clinical decision making? • not at all useful • somewhat useful • very useful	The number of response anchors influences the reliability of a set of survey items.[5] Using too few response anchors generally reduces reliability. There is, however, a point of diminishing returns beyond which more response anchors do not enhance reliability.[5]	Use five or more response anchors to achieve stable participant responses. In most cases, using more than seven to nine anchors is unlikely to be meaningful to most respondents and will not improve reliability.[5]	How useful was your medical school training in clinical decision making? • not at all useful • slightly useful • moderately useful • quite useful • extremely useful

References:
1. Dillman DA, Smyth JD, Christian LM. Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method. 3rd ed. New York, NY: John Wiley & Sons; 2009.
2. Schwarz N. Self-reports: How the questions shape the answers. Am Psychol. 1999;54:93–105.
3. Krosnick JA. Survey research. Annu Rev Psychol. 1999;50:537–567.
4. Tourangeau R, Rips LJ, Rasinski KA. The Psychology of Survey Response. New York, NY: Cambridge University Press; 2000.
5. Weng L. Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educ Psychol Meas. 2004;64:956–972.

Disclaimers:
The views expressed in this article are those of the authors and do not necessarily reflect the official policy of the Department of Defense.
Dr. Steven Durning coauthored this Last Page prior to becoming assistant editor, AM Last Page.

AM Last Page: Avoiding Four Visual-Design Pitfalls in Survey Development

Anthony R. Artino, Jr., PhD, associate professor, Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, and Hunter Gehlbach, PhD, associate professor, Harvard Graduate School of Education

A previous AM Last Page[1] presented five common pitfalls of survey design as well as several solutions. This AM Last Page presents four visual-design and layout pitfalls and offers solutions.

Pitfall: Explanation and Example	Solution: Explanation and Example
Labeling only the end points of your response options Labeling only the end points leaves the meaning of the unlabeled options open to respondents' interpretation. Different respondents can interpret the unlabeled options differently. This ambiguity increases measurement error.[2] How interesting did you find this clinical reasoning course? ○ ○ ○ ○ ○ not at all interesting … extremely interesting	**Verbally label each response option** Labeling each response option increases consistency in the conceptual spacing between response options and increases the likelihood that all respondents will interpret the response options similarly. Additionally, the visual weight of the response options are comparable, so that respondents' eyes are not drawn to certain options. How interesting did you find this clinical reasoning course? ○ not at all interesting ○ slightly interesting ○ moderately interesting ○ quite interesting ○ extremely interesting
Labeling response options with both numbers and verbal labels Because of the additional information respondents must process, providing both numbers and verbal labels extends response time.[3] The implied meaning of negative numbers can be particularly confusing, and may introduce additional error. For example, in the item below, learning "a little bit" seems incongruous with learning the amount of "-1." How much did you learn in today's workshop? ○ -2 almost nothing ○ -1 a little bit ○ 0 some ○ 1 quite a bit ○ 2 a tremendous amount	**Use only verbal labels** In general, use only verbal labels for each response option. Doing so will reduce the cognitive effort required of your respondents and will likely reduce measurement error.[2] How much did you learn in today's workshop? ○ almost nothing ○ a little bit ○ some ○ quite a bit ○ a tremendous amount
Unequally spacing your response options The visual spacing between options can attract respondents to certain options over others, which in turn might cause them to select these options more frequently.[4] In addition, unbalanced spacing of the response options can shift the visual midpoint of the scale. How much did you learn from your peers in this course? ○ almost nothing ○ a little bit ○ some ○ quite a bit ○ a tremendous amount	**Maintain equal spacing between response options** Maintaining equal spacing between response options will reinforce the notion that, conceptually, there is equal space or "distance" between each response option. As a result, the answers will be less biased, thereby reducing measurement error. How much did you learn from your peers in this course? ○ almost nothing ○ a little bit ○ some ○ quite a bit ○ a tremendous amount
Placing non-substantive response options together with substantive response options Placing non-substantive response options such as "don't know," "no opinion," or "not applicable" together with the substantive options can shift the visual and conceptual midpoint of the response scales, thereby skewing the results.[4] How satisfied are you with the quality of the library services? ○ not at all satisfied ○ slightly satisfied ○ moderately satisfied ○ quite satisfied ○ extremely satisfied ○ not applicable	**Use additional space to visually separate non-substantive response options** Using additional space to visually separate non-substantive response options from the substantive options will align the visual midpoint with the conceptual midpoint thereby reducing measurement error.[4] This recommendation is an important exception to the guidance above about maintaining equal spacing between response options. How satisfied are you with the quality of the library services? ○ not at all satisfied ○ slightly satisfied ○ moderately satisfied ○ quite satisfied ○ extremely satisfied ○ not applicable

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

References:

1. Artino AR Jr., Gehlbach H, Durning SJ. AM Last Page: Avoiding five common pitfalls of survey design. Acad Med. 2011;86:1327.
2. Krosnick JA. Survey research. Annu Rev Psychol. 1999;50:537-567. http://communication.stanford.edu/faculty/krosnick/docs/annrevsurvey.pdf. Accessed May 30, 2012.
3. Christian LM, Parsons NL, Dillman DA. Designing scalar questions for web surveys. Sociol Meth Res. 2009;37:393-425.
4. Dillman DA, Smyth JD, Christian LM. Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2009.

Author contact: Anthony.Artino@usuhs.mil

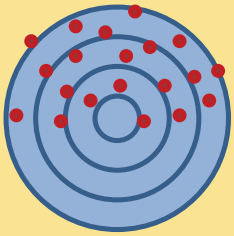
AM Last Page: Reliability and Validity in Educational Measurement

Anthony R. Artino, Jr., PhD, assistant professor of preventive medicine and biometrics, Steven J. Durning, MD, professor of medicine, and Alisha H. Creel, PhD, assistant professor of social and behavioral sciences, Uniformed Services University of the Health Sciences

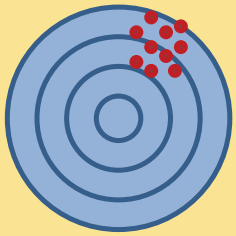
Reliability is the extent to which the scores produced by a particular measurement tool or procedure are consistent and reproducible.[1] Reliability answers the question, "Does the assessment yield the same scores at different times, from different raters, or from different items?"

Validity is the degree to which an assessment measures what investigators want to measure, all of what they want to measure, and nothing but what they want to measure.[1] Validity answers the question, "Does the assessment provide information that is relevant to the inferences that are being made from it?" An assessment, such as a test or questionnaire, does not have validity in any absolute sense. Instead, the scores produced are valid for some uses and not valid for others.

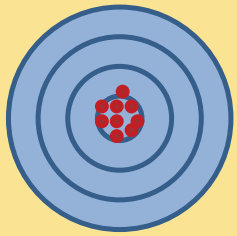
A target provides a metaphor for the relationship between reliability and validity.
The true score (or value) for the concept the researcher is attempting to measure is at the center of the target, and the observed score the investigator gets from each person assessed is a shot at the target.



Neither reliable nor valid



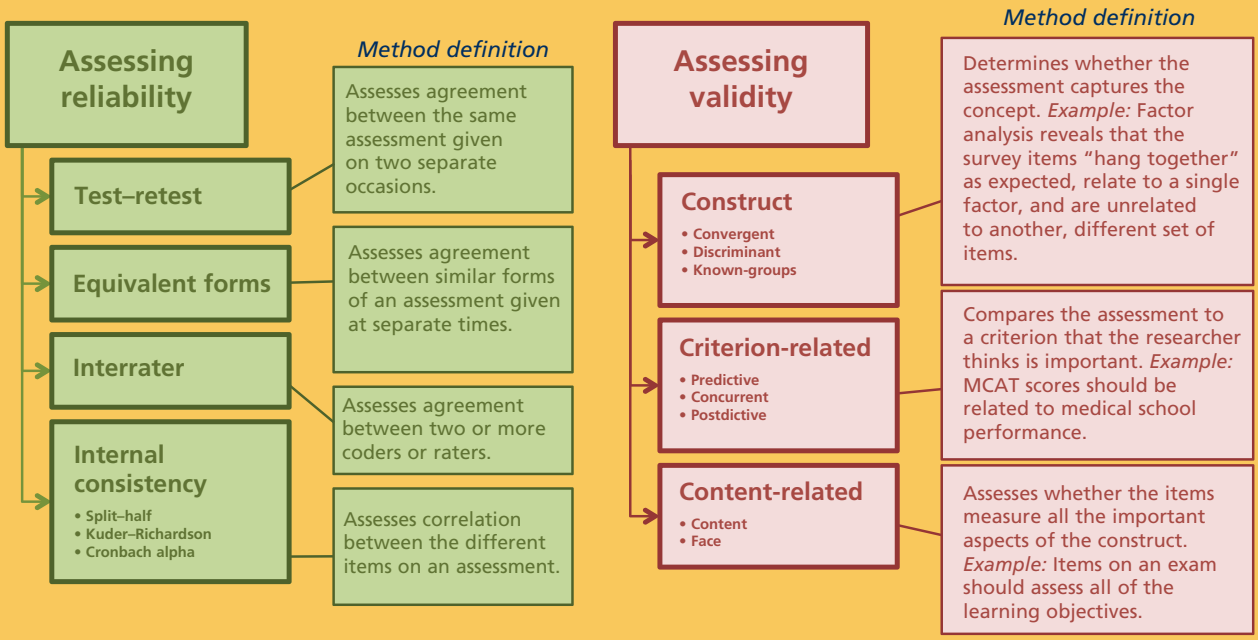
Reliable, but not valid



Both reliable and valid

Reliability is a *necessary but insufficient* condition for validity. To be valid, scores must first be at least moderately reliable.[1–3] However, scores that are reliable may be devoid of validity for the application the researcher has in mind.[1]

Many methods of assessing reliability and validity are available.[1–4] Each method provides the researcher with slightly different information about the reliability and validity of the assessment.



References

1. Thorndike RM. Measurement and Evaluation in Psychology and Education. 7th ed. Upper Saddle River, NJ: Pearson Education; 2005.
2. 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.
3. Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ. 2003;37:830–837.
4. Kane MT. Validation. In: Educational Measurement, 4th ed. Brennan RL, ed. Westport, CT: American Council on Education and Praeger Publishers; 2006:17–64.

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 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

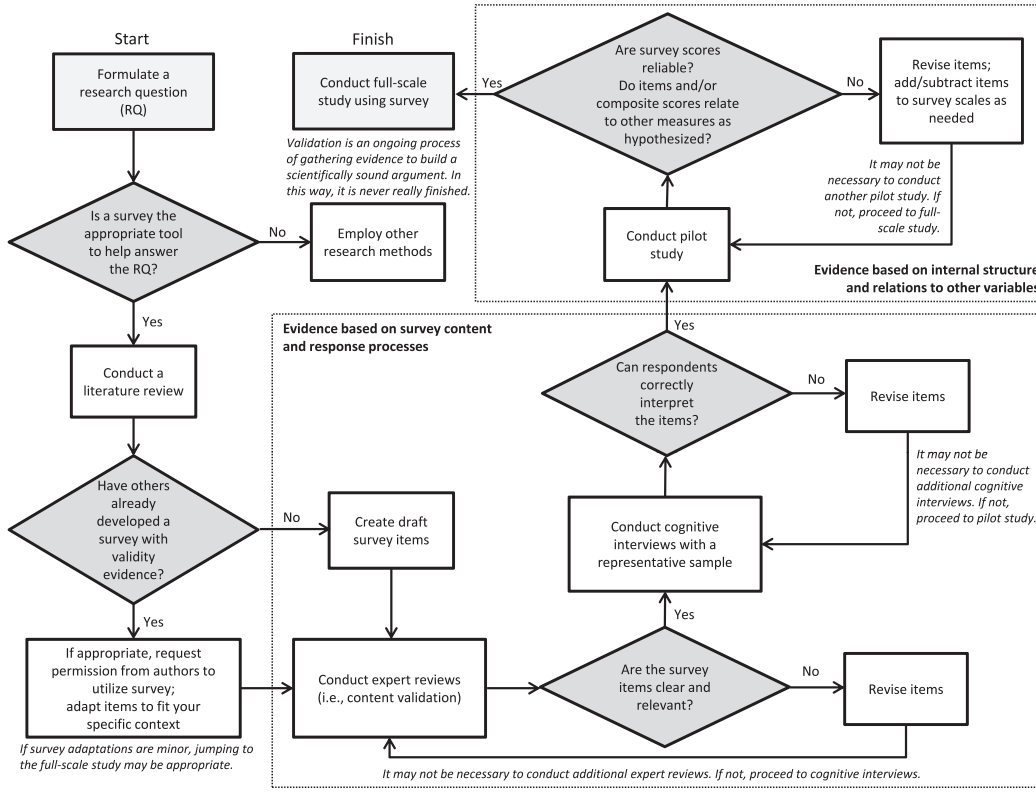


FIGURE **A Systematic Approach to Survey Design for Graduate Medical Education Research**

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.

What Do Our Respondents Think We're Asking? Using Cognitive Interviewing to Improve Medical Education Surveys

Gordon B. Willis, PhD
Anthony R. Artino Jr., PhD

Consider the last time you answered a questionnaire. Did it contain questions that were vague or hard to understand? If yes, did you answer these questions anyway, unsure if your interpretation aligned with what the survey developer was thinking? By the time you finished the survey, you were probably annoyed by the unclear nature of the task you had just completed. If any of this sounds familiar, you are not alone, as these types of communication failures are commonplace in questionnaires.[1–3] And if you consider how often questionnaires are used in medical education for evaluation and educational research, it is clear that the problems described above have important implications for the field. Fortunately, confusing survey questions can be avoided when survey developers use established survey design procedures.

In 2 recent *Journal of Graduate Medical Education* editorials,[4,5] the authors encouraged graduate medical education (GME) educators and researchers to use more systematic and rigorous survey design processes. Specifically, the authors proposed a 6-step decision process for questionnaire designers to use. In this article, we expand on that effort by considering the fifth of the 6 decision steps, specifically, the following question: "Will my respondents interpret my items in the manner that I intended?" To address this question, we describe in detail a critical, yet largely unfamiliar, step in the survey design process: cognitive interviewing.

Questionnaires are regularly used to investigate topics in medical education research, and it may seem a straightforward process to script standardized survey questions. However, a large body of evidence demonstrates that items the researchers thought to be perfectly clear are often subject to significant misinterpretation, or otherwise fail to measure what was intended.[1,2] For instance, abstract terms like "health professional" tend to conjure up a wide range of interpretations that may depart markedly from those the questionnaire designer had in mind. In this example, survey respondents may choose to include or exclude marriage counselors, yoga instructors, dental hygienists, medical office receptionists, and so on, in their own conceptions of "health professional." At the same time, terms that are precise but technical in nature can produce unintended interpretations; for example, a survey question about "receiving a dental sealant" could be misinterpreted by a survey respondent as "getting a filling."[2]

The method we describe here, termed "cognitive interviewing" or "cognitive testing," is an evidence-based, qualitative method specifically designed to investigate whether a survey question—whether attitudinal, behavioral, or factual in nature—fulfills its intended purpose (BOX). The method relies on interviews with individuals who are specifically recruited. These individuals are presented with survey questions in much the same way as survey respondents will be administered the final draft of the questionnaire. Cognitive interviews are conducted before data collection (pretesting), during data collection, or even after the survey has been administered, as a quality assurance procedure.

During the 1980s, cognitive interviewing grew out of the field of experimental psychology; common definitions of cognitive interviewing reflect those origins and emphasis. For example, Willis[6] states, "Cognitive interviewing is a psychologically oriented method for empirically studying the way in which individuals mentally process and respond to survey questionnaires." For its theoretical underpinning, cognitive interviewing has traditionally relied upon the 4-stage cognitive model introduced by Tourangeau.[7] This model describes the survey response process as involving (1) comprehension, (2) retrieval of information, (3) judgment or estimation, and (4) selection of a response to the question. For example, mental processing of the question "In the past year, how many times have you participated in a formal educational program?" presumably requires a respondent to comprehend and interpret critical terms and phrases (eg, "in the past year" and "formal educational program"); to recall the correct answer; to decide to report an accurate number (rather

Gordon B. Willis, PhD, is Cognitive Psychologist, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health; and **Anthony R. Artino Jr, PhD,** is Associate Professor of Medicine and Preventive Medicine & Biometrics, Uniformed Services University of the Health Sciences.

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

DOI: http://dx.doi.org/10.4300/JGME-D-13-00154.1

BOX **Glossary**

Cognitive interviewing (or cognitive testing)—an evidence-based qualitative method specifically designed to investigate whether a survey question satisfies its intended purpose.
Concurrent probing—a verbal probing technique wherein the interviewer administers the probe question immediately after the respondent has read aloud and answered each survey item.
Proactive verbal probes—a verbal probe that is systematically fashioned before the cognitive interview in order to search for potential problems.
Reactive verbal probes—a verbal probe that is developed in a nonstandardized way during the cognitive interview, often in response to a respondent's behavior.
Reactivity effects—bias in a respondent's behavior caused by the additional cognitive demands related to answering probe questions and explaining those answers (most likely by leading to more carefully thought-out answers than a survey respondent might typically provide).
Verbal probing—a cognitive interviewing technique wherein the interviewer administers a series of probe questions specifically designed to elicit detailed information beyond that normally provided by respondents.
Retrospective probing—a verbal probing technique wherein the interviewer administers the probe questions after the respondent has completed the entire survey (or a portion of the survey).
Think-aloud interviewing—a cognitive interviewing technique wherein survey respondents are asked to actively verbalize their thoughts as they attempt to answer the evaluated survey items.

than, for example, providing a higher value); and then to produce an answer that matches the survey requirements (eg, reporting "5 times" rather than "frequently"). Most often, comprehension problems dominate. For example, it may be found that the term "formal educational program" is variably interpreted. In other words, respondents may be unsure which activities to count and, furthermore, may not know what type of participation is being asked about (eg, participation as a student, teacher, or administrator).

More recently, cognitive interviewing has to some extent been reconceptualized as a sociological/anthropological endeavor, in that it emphasizes not only the individualistic mental processing of survey items but also the background social context that may influence how well questions meaningfully capture the life of the respondent.[8] Especially as surveys increasingly reflect a range of environments and cultures that may differ widely, this viewpoint has become increasingly popular. From this perspective, it is worth considering that the nature of medical education may vary across countries and medical systems, such that the definition of a term as seemingly simple as "graduate medical education" might itself lack uniformity.

Process Overview

Cognitive interviewing is conducted using 2 key procedures that characterize the method and that require more than the usual practice of a respondent simply answering a presented survey question. First, *think-aloud interviewing* requests that the survey respondents (often renamed subjects or participants) actively verbalize their thoughts as they attempt to answer the survey questions.[9] Here the role of the interviewer is mainly to support this activity by asking the subject to "keep talking" and to record the resultant verbal record, or so-called cognitive protocol, for later analysis. For this procedure to be effective, it is necessary that the subject's think-aloud stream contain diagnostic information relevant to the assessment of survey question function, as opposed to tangential or meandering free association. The alternative procedure, which has gained considerable traction over the past 20 years, is *verbal probing*, which is a more active form of data collection in which the cognitive interviewer administers a series of probe questions specifically designed to elicit detailed information beyond that normally provided by respondents. Commonly used probes are provided in TABLE 1.

Probes may be systematically fashioned before the interview, in order to search for potential problems (proactive probes), or they may be developed in a nonstandardized way during the interview, often in response to subject behavior (reactive probes). As such, probing is usually a flexible procedure that relies to a significant degree on interviewer adaptability rather than on strict standardization of materials.

The advantages to a think-aloud procedure is that it presents little in the way of training demands on the interviewer and avoids being too directive in ways that could bias responses. That being said, thinking aloud places a significant burden on subjects, many of whom find the process to be unnatural and difficult. Further, analysis of think-aloud transcripts can be a daunting task because of the sheer quantity of the sometimes meandering verbalizations obtained. Verbal probing, on the other hand, requires more training and thought on the part of the interviewer and, if not done carefully, can create reactivity effects. That is, verbal probing may create bias in the subject's behavior through the additional cognitive demands related to answering probe questions and explaining those answers. For example, probing may lead to more carefully thought-out answers than a survey respondent might typically provide. However, verbal probes are generally efficient, can be targeted toward specific cognitive processes (eg, comprehension of a term or phrase), are well-accepted by subjects, and provide data that are easier to separate into segments and code than think-aloud streams. In practice, think-aloud interviewing and verbal probing are very often used in unison, with a distinct emphasis on the latter. TABLE 2 provides several examples of survey items that were tested with verbal probes during cognitive interviews, revealing varying interpretations.

Analysis of Cognitive Interviews

Because cognitive interviewing is a qualitative procedure, analysis does not rely on strict statistical analysis of

TABLE 1 **Categories and Examples of Cognitive Probe Questions (Adapted from Willis[2])**

Type of Cognitive Probe	Example
Comprehension/interpretation	"What does the term 'formal educational program' mean to you?"
Paraphrasing	"Can you repeat the question I just asked in your own words?"
Confidence judgment	"How sure are you that you have participated in 5 formal educational programs?"
Recall	"How do you remember that you have participated in 5 formal educational programs?"
	"How did you come up with your answer?"
Specific	"Why do you say that you think it is very important that physicians participant in continuing education?"
General	"How did you arrive at that answer?"
	"Was that easy or hard to answer?"
	"I noticed that you hesitated. Tell me what you were thinking."
	"Tell me more about that."

numeric data but rather on coding and interpretation of written notes that are taken during the interview (often by the interviewer) or after it (by either the interviewer or an analyst). Such notes often describe substantive observations relevant to item functioning (eg, "Subject was unsure of what we mean by 'formal educational program,' in terms of what counts and what should be excluded"). A set of interviews are typically analyzed by combining the notes pertaining to each evaluated item, aggregating across interviews, seeking common themes, and identifying key findings that may indicate departures from the interpretation desired by survey developers. For example, if one finds that there is little concurrence across a range of subjects about what is meant by "formal educational program," a potential solution may be to explicitly define this term in the survey. Furthermore, if clear patterns are observed where the interpretation of an item is found to be culturally specific, or varies across country or context, the researcher may find it necessary to revisit whether the question can in fact be asked in the same form across these contexts.

Practical Considerations

Cognitive interviewing projects are normally small in scope and may involve just 10 to 30 total subjects. For small-scale GME projects, as few as 5 or 6 subjects may provide useful information to improve survey items, as long as the survey developers are sensitive to the potential for bias in very small samples. Cognitive interviewing is often conducted as an

TABLE 2 **Questionnaire Items Tested via Cognitive Interviewing and Found to Be Confusing or Misinterpreted (Adapted from Willis[2])**

Question Tested	Probe Question	Subject Response
1. Are you now bothered by pain? *Note: The purpose of this item was to obtain a measure of pain experience.*	"Why did you say 'no'?"	"Well, yeah, it hurts, but the Marines taught us to be tough, so I try not to let it bother me. So I said no."
2. In the past 12 months, have you received any complementary medicine? *Note: The purpose of this item was to obtain estimates of use of complementary and alternative medicine.*	"What, to you, is 'complementary medicine'?"	"That would be medical care that's free."
3. Do you think that genes influence someone's weight? *Note: This item was part of a larger scale intended to determine attitudes concerning genetic influences on cancer risk factors.*	"Why did you say 'yes'?"	"Because if you wear tight jeans, you will want to eat less so they will fit."
4. In the last year, how often did you experience pain? *Note: This item was meant to assess pain experienced over the past 365 days.*	"What does 'in the last year' mean to you?"	Subject 1: "This calendar year."
		Subject 2: "Last calendar year."
		Subject 3: "The last 365 days."

iterative procedure, in which a small round of interviews is conducted (eg, 10); the survey developers analyze and assess their results, make alterations to survey items where necessary, and then conduct an additional round of testing using the modified items. In the previous example, 3 rounds of 10 interviews would likely provide ample opportunity for problematic items to emerge. Such items could then be reevaluated by the survey developer and modified if needed.

Cognitive interviewing is commonly applied to both interviewer-administered and self-administered surveys and for surveys designed for paper- and computer-based administration. Under interviewer administration, it is common to rely on concurrent probing, which consists of probing after each evaluated question is read aloud and answered. The back and forth process of presenting a question, having the subject answer it, and then asking targeted probes appears to function well for interviewer-administered surveys and poses little difficulty for most people. Under self-administration, where the interviewer is normally silent or absent and the respondent completes the survey unaided, the process of retrospective probing is often applied. In retrospective probing, interviewer questions are reserved for a debriefing session that occurs after the subject has completed the questionnaire. Concurrent or retrospective probing can be applied to either interviewer or self-administered surveys, and there are advantages as well as limitations to each. Concurrent probing allows the subject to respond to probes when their thoughts are recent and presumably fresh, whereas retrospective probing requires revisiting thoughts that are more remote in time. However, concurrent probing can disrupt the interview through its imposition of probes in a way that retrospective probing does not.

With proper training, cognitive interviews can be carried out by anyone planning to administer a questionnaire. Because the equipment and logistical requirements are modest—all that is really needed is a quiet place and an audio recorder—it is possible for a wide range of educators or researchers to conduct the activity with subjects who are similar to the target survey population. Usually the objective of recruitment is not to achieve any type of statistical representation but rather to cover the territory by including as wide a range of subjects as possible, given the constraints of time and cost. It is important to note, however, that data obtained from cognitive interviews are qualitative in nature and used to assess and improve survey items, usually before the survey is implemented. As such, these data and the respondents used for the cognitive interviews should not be used as part of the final survey study.

Concluding Thoughts

Creating a survey with evidence of reliability and validity can be difficult. The cognitive interviewing processes outlined here are designed to help GME educators and researchers improve the quality of their surveys and enhance the validity of the conclusions they draw from survey data. Cognitive interviewing is an evidence-based tool that can help survey developers collect validity evidence based on survey content and the thought processes that participants engage in while answering survey questions.[2,8–10] As with all of the steps in survey design, the results of cognitive interviewing should be reported in the methods section of research papers. Doing so gives readers greater confidence in the quality of the survey information reported by GME researchers.

References

1 Tourangeau R, Rips LJ, Rasinski K. *The Psychology of Survey Response*. New York, NY: Cambridge University Press; 2000.
2 Willis GB. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications; 2005.
3 Willis GB. *Cognitive Interviewing: A "How-To" Guide*. 1999. http://appliedresearch.cancer.gov/areas/cognitive/interview.pdf. Accessed May 22, 2013.
4 Magee C, Byars L, Rickards G, Artino AR Jr. Tracing the steps of survey design: a graduate medical education research example. *J Grad Med Educ*. 2013;5(1):1–5.
5 Rickards G, Magee C, Artino AR Jr. You can't fix by analysis what you've spoiled by design: developing survey instruments and collecting validity evidence. *J Grad Med Educ*. 2012;4(4):407–410.
6 Willis G. Cognitive aspects of survey methodology. In: Lavrakas P, ed. *Encyclopedia of Survey Research Methods*. Vol 2, Thousand Oaks, CA: Sage Publications; 2009:103–106.
7 Tourangeau R. Cognitive science and survey methods: a cognitive perspective. In: Jabine T, Straf M, Tanur J, Tourangeau R, eds. *Cognitive Aspects of Survey Design: Building a Bridge between Disciplines*. Washington, DC: National Academy Press; 1984:73–100.
8 Gerber ER. The view from anthropology: ethnography and the cognitive interview. In: Sirken M, Herrmann D, Schechter S, Schwarz N, Tanur J, Tourangeau R, eds. *Cognition and Survey Research*. New York, NY: Wiley; 1999:217–234.
9 Ericsson KA, Simon HA. Verbal reports as data. *Psychol Rev*. 1980;87:215–251.
10 Conrad F, Blair J. Data quality in cognitive interviews: the case for verbal reports. In: Presser S, Rothgeb J, Couper M, Lessler J, Martin E, Martin J, Singer E, eds. *Questionnaire Development Evaluation and Testing Methods*. Hoboken, NJ: John Wiley and Sons; 2004:67–87.

Survey Design

Glossary of Terms

Acquiescence is the tendency to endorse any assertion made in a question, regardless of its content (this is really a type of satisficing; see definition below).

Construct is a hypothesized concept, model, idea, or theory (something constructed) that we think exists but that we cannot directly observe.

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

Factor is an "unobserved" variable that statistically explains the variation and co-variation among a larger set of "observed" variables (i.e., the actual items on a survey). Stated another way, factors succinctly represent a larger set of observed variables. Factors often correspond to constructs; although some constructs are made up of multiple factors. Such constructs are often called multi-dimensional constructs.

Factor analysis is an analytical technique used to identify factors that statistically explain the variation and co-variation among a set of measures (i.e., a set of survey items). Factor analysis is a data-reduction technique that reduces a large number of overlapping measured variables to a much smaller set of factors.

Items/Indicators (observable items, empirical indicators) are the actual questions/statements that make up a survey (or a particular survey scale).

Optimizing is the extent to which a respondent performs the necessary cognitive tasks to answer a survey item in a thorough and unbiased manner. These cognitive tasks may include: (1) interpreting a survey item (figuring out its intent), (2) searching memory for relevant information, (3) forming a judgment, and (4) translating the judgment into an answer by summarizing or selecting one of the alternatives offered. These are the tasks we *want* respondents to do when taking our survey.

Order effect is the notion that the order of response alternatives affects the extent to which respondents select those items (primary and recency effects are two types of order effects).

Primacy effect is the tendency to remember (and select) answers that appear first (or early) in a list of alternatives (likely because those items were cognitively processes and now reside in long-term member). This effect is more prominent when items are presented visually.

Recency effect is the tendency to remember (and select) answers that appear last (or later) in a list of alternatives (likely because they still reside in working memory and so are more accessible). This effect is more prominent when items are presented orally.

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

Response anchors are the named points along a set of answer options (e.g., strongly disagree, disagree, neutral, agree, strongly agree).

Survey Design

Glossary of Terms

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

Satisficing is the extent to which respondents compromise their standards and expend less energy (i.e., they don't fully optimize).

Scale is two or more items (indicators) intended to measure a construct. Often, however, the word scale is used more generally to refer to the entire survey. As such, many scales are composed of several sub-scales.

Social desirability bias is the tendency to over-report admirable attitudes/behaviors and under-report those that are not socially respected. Stated another way, it is the tendency to lie in order to appear as socially suitable and acceptable as possible.

Strong satisficing is a more dramatic form of satisficing where respondents skip entire cognitive tasks (i.e., comprehension, retrieval, judgment, or response selection) and arbitrarily select an answer (e.g., they may select the first reasonable response; they may accept any assertions made that seem reasonable; they may select "don't know" or "no opinion" to avoid expending effort; they may randomly select a response from those offered).

Sub-scale is a sub-division of a larger scale. Often, multi-dimensional constructs will be measured with a scale that is made up of several smaller sub-scales.

Weak satisficing is a less serious form of satisficing where respondents are less thorough in comprehension, retrieval, judgment, and response selection (e.g., they may be less thoughtful about a question's meaning; they may search their memories less comprehensively; they may integrate retrieved information carelessly; they may select a response imprecisely).

Validity is the degree to which evidence and theory support a measure's intended use.

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

Construct-Specific Response Scales

					Construct:
not important	somewhat important	important	very important	extremely important	importance
unimportant	of little importance	moderately important	important	very important	importance
not at all important	slightly important	moderately important	quite important	extremely important	importance
completely unimportant	unimportant	neutral	important	completely important	importance
not at all confident	slightly confident	moderately confident	quite confident	extremely confident	self-efficacy (confidence)
completely dissatisfied	moderately dissatisfied	neutral	moderately satisfied	completely satisfied	satisfaction
not at all satisfied	slightly satisfied	moderately satisfied	quite satisfied	extremely satisfied	satisfaction
not at all bored	slightly bored	moderately bored	quite bored	extremely bored	boredom
not at all frustrated	slightly frustrated	moderately frustrated	quite frustrated	extremely frustrated	frustration
strongly prefer x	prefer x	neutral	prefer y	strongly prefer y	comparing x to y
almost no effort	a little bit of effort	some effort	quite a bit of effort	a great deal of effort	effort
very poor	poor	barely acceptable	good	very good	quality

More General Response Scales

completely untrue	somewhat untrue	yes and no	somewhat true	completely true	
not at all true of me	slightly true of me	somewhat true of me	mostly true of me	completely true of me	
strongly disagree	disagree	neutral	agree	strongly agree	
		or neither agree nor disagree			
disagree strongly	disagree	tend to disagree	tend to agree	agree	agree strongly
disagree strongly	disagree moderately	disagree slightly	agree slightly	agree moderately	agree strongly
disagree very strongly	disagree strongly	disagree	agree	agree strongly	agree very strongly
completely disagree	disagree	neutral	agree	completely agree	
completely disagree	mostly disagree	slightly disagree	slightly agree	mostly agree	completely agree

Frequency or "Degree" Response Scales

almost never	once in a while	sometimes	often	almost all the time	
never	seldom	about half the time	usually	always	
never	little	somewhat	much	a great deal	
not at all	very little	moderately	quite a bit	a tremendous amount	
never	rarely	occasionally	frequently	almost always	
never	seldom	sometimes	often		
never	rarely	sometimes	often	very often	always
seldom	occasionally	to a considerable degree	almost always		
never	very rarely	rarely	occasionally	frequently	very frequently
never	very rarely	rarely	occasionally	very frequently	always

Survey Development References

Good General Textbooks & Articles:

Bradburn, N. M., Rips, L, J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science, 236,* 157-161.

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

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). New York: John Wiley & Sons.

Fowler, F. J. (2002). *Survey research methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537-567.

Schwarz, N. (1999). Self-reports. How the questions shape the answers. *American Psychologist, 54,* 93-105.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. New York: Cambridge University Press.

Articles on Expert Validation:

McKenzie, J.F., Wood, M.L., Kotecki, J.E., Clark, J.K., & Brey, R.A. (1999). Establishing content validity: Using qualitative and quantitative steps. *American Journal of Health Behavior, 23,* 311-318.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27,* 94-104.

Articles on Cognitive Interviewing:

Karabenick, S.A., Woolley, M.E., Friedel, J.M., Ammon, B.V., Blazevski, J., Bonney, C.R., et al. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 4*2, 139-151.

Willis, G.B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage Publications.

Articles on Reliability and Factor Analyses:

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

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,* 272-299.

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,* 393-416.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350-353.

Thompson, B. (1994). Guidelines for authors reporting score reliability estimates. *Educational and Psychological Measurement, 54,* 837-847.