Tracing the Steps of Survey Design: A Graduate Medical Education Research Example

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

Surveys are frequently used to collect data in graduate medical education (GME) settings.[1] However, if a GME survey is not rigorously designed, the quality of the results is likely to be lower than desirable. In a recent editorial we introduced a framework for developing survey instruments.[1] This systematic approach is intended to improve the quality of GME surveys and increase the likelihood of collecting survey data with evidence of reliability and validity. In this article we illustrate how researchers in medical education may operationalize this framework with examples from a survey we developed during the recent integration of 2 independent internal medicine (IM) residency programs.

Background

In 2010, the Department of Defense mandated the integration of the Walter Reed Army Medical Center in Washington, DC, with the National Naval Medical Center in Bethesda, Maryland. Prior to this integration, each hospital maintained independently accredited GME programs, including separate IM residency programs. During the merger these IM programs were asked to integrate seamlessly into a unified program.

Despite many similarities, the 2 IM programs had important differences that might inhibit successful integration. For example, residents at Walter Reed were accustomed to an overnight, 24-hour call structure that was thought to bring a strong experiential learning element to the program. Yet, this call system risked violating work hour restrictions. Residents at the National Naval Medical Center worked under a night-float system that eliminated the risk of duty hour violations but increased the number of handoffs. Given these programmatic differences, we were interested in understanding how individuals in both programs thought the integration would affect the quality of the IM residency.

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

The authors are military service members. The views expressed in this article are those of the authors and do not necessarily reflect the official policy of the US Department of Defense.

Corresponding author: Anthony R. Artino Jr, PhD, Department of Preventive Medicine and Biometrics, 4301 Jones Bridge Road, Bethesda, MD 20814, 301.295.3693, anthony.artino@usuhs.edu

DOI: http://dx.doi.org/10.4300/JGME-D-12-00364.1

Our Survey Design Process

Evidence-based design processes allow GME researchers to develop a set of survey items that every respondent is likely to interpret the same way, is able to respond to accurately, and is willing and motivated to answer. Six questions, introduced in our first editorial, can guide researchers through this systematic survey design framework. In the sections that follow, we illustrate each step of the process with examples from our own survey design project. Readers interested in the rationale behind the framework or in further details about each step are encouraged to consult our previous article.[1]

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

Before creating a survey, it is important to consider the research question(s) of interest and the variables (or constructs) the researcher intends to measure. If, for example, the research question relates to the beliefs, opinions, or attitudes of the intended audience, a survey makes sense. On the other hand, if a researcher is more interested in assessing a directly observable behavior, such as residents' skill level for a particular clinical procedure, an observational tool may be a better choice.

In the context of the residency merger, we wanted to understand how the integration effort would have an impact on key GME quality elements and program requirements as specified by the Accreditation Council for Graduate Medical Education (ACGME).[2] We believed that understanding these factors from the residents' perspective would enable leadership to identify threats to successful integration as well as potential opportunities for process improvement. Further, we felt a survey was the appropriate tool because it would allow us to collect real-time feedback from participants rather than waiting for more objective outcomes, such as in-service exam scores and board scores, which, although valuable, would occur much later. The residents of each program were identified as the target population for our survey.

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

A thorough review of the literature should be the next step in the GME survey design process. This step provides

information about how the construct of interest has been defined in previous research. It also helps one identify existing survey scales that could be employed or adapted.

Our review revealed that very little has been published on integration of GME programs. However, we found a number of examples of organizational change and restructuring in the business literature. Some of the most widely published, well-studied examples of organizational change were developed by William Bridges.[3] We reviewed several of Bridges' survey instruments to identify common themes and items that might be applicable to our survey. Ultimately, we did not use any of these items verbatim. Instead, we adapted several items and used Bridges' work to better define our constructs of 3 separate but related ideas: current satisfaction, perceptions of the impact of the integration on the training experience, and beliefs about the readiness of the training programs to make the transition.

GME researchers who find and wish to use or modify relevant survey scales can usually contact authors and request such use. It is worth noting, however, that "previously validated" survey scales require the collection of additional reliability and validity evidence in the specific research context, particularly if the scales are modified in any way or used in populations different from the initial survey audience. For publication, this additional evidence should be reported in the "Methods" and "Results" sections.

Question 3: How Do I Develop My Survey Items?

The goal of this step is to create survey items that adequately represent the construct of interest in a language that respondents can easily understand. One important design consideration is the number of items needed to adequately assess the construct. There is no easy answer to this question. The ideal number of items depends on a number of factors, including the complexity of the construct and the level at which one intends to assess the construct (sometimes referred to as the "grain size" or level of abstraction at which the construct will be measured).[4] In general, it is a good idea to develop more items than will ultimately be needed in the final scale because some items will undoubtedly be deleted or revised later in the design process.[4]

The next challenge is to write a set of clear and unambiguous items. Writing good items is as much an art as it is a science. Nonetheless, there is a plethora of item-writing guidance—evidence-based, best practices—that should be used to guide the item-writing process.[1,4–8] Reviewing these best practices is beyond the scope of this editorial; however, we have provided a summary of several evidence-based recommendations in TABLE 1 to assist readers.

To guide our item-writing process, we selected elements that the ACGME requires in an accredited IM residency.[2] In particular, we chose elements that we felt were likely to be affected by reorganization and were also visible to the residents. Our initial draft had questions about every potentially relevant element; it quickly became clear that we had too many items. As such, we refocused our efforts on those issues most likely to be relevant to the intended respondents. In doing so, we were able to cut down our survey from 150 items to a more manageable 45 items.

To illustrate other decisions we made during the survey development process, we focus the remainder of this editorial on our *didactic quality* scale. For this scale, we wanted to know the extent to which participants believed the merger would have an impact on the quality of their IM residency experience. For example, one item asked "How do you think the internal medicine integration will impact the educational quality of morning report?" Because we believed, based on our literature review and discussion with experts, that respondents might think the impact could either be positive or negative, we chose a bipolar scale with a midpoint of "neither positive nor negative impact" and endpoints of "extreme negative impact" on the low side of the response scale and "extreme positive impact" on the high side of the response scale (TABLE 2).

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

To assess survey content, GME researchers should ask a group of experts to review the items. This process, called content or expert validation, involves asking experts to review the draft survey items for clarity, relevance to the construct, and cognitive difficulty.[9–11] Experts can also assist in identifying important aspects of the construct that may have been omitted during item development. "Experts" might include those more experienced in survey design, national content experts, or local colleagues knowledgeable about the specific construct of interest. The number of experts needed to conduct a content validation is typically small; 6 to 12 experts will often suffice.[10,11]

Our 9 experts included staff from each IM program, as well as select faculty from our affiliated university who had expertise in survey design. Each expert received an invitation to participate in the content validation, along with the draft survey items and a document outlining our purpose and the specific aspects of the survey on which we wanted him or her to focus. Through this process our experts identified 6 items that were poorly focused; we eliminated these items from the survey. Our experts did not identify any content omissions in the scale items, and they agreed with our use of ACGME quality elements and program requirements as universal attributes of a high-quality GME training program. Finally, our experts identified several items that they felt were difficult to interpret, and these items were revised.

TABLE 1 **Evidence-Based Best Practices for Writing Survey Items**

Frequently Asked Questions	Best Practice	Rationale
1. Should I write my survey items in the form of a question or a statement?	Where possible, write survey items in the form of a question rather than a statement.[6,8]	Questions are more conversational and respondents are more practiced at answering questions as opposed to rating a set of statements.
2. What type of response options should I use, agreement response options or some other type?	Avoid agreement response options. Instead, use construct-specific response options where possible.[6,8]	Agreement response options do not emphasize the construct being assessed and may encourage respondents to acquiesce; that is, to agree with the item regardless of its content.
3. How many response options should I use?	Use at least 5 response options and no more than 9.[6,8]	Using too few response options tends to reduce the reliability of a set of survey items. Providing too many response options will not be meaningful to most respondents. Also, using too many response options can give the false impression of high precision yet is unlikely to improve reliability.
4. Should I use an odd or an even number of response options?	Use an odd number of response options if your construct has a conceptual midpoint.[8]	Although there is no definitive answer to the question of whether to use an odd or an even number of response options, in many cases having a midpoint can encourage accuracy. This is particularly true if your construct has a conceptual midpoint (eg, a neutral point), which many constructs do.
5. How should I label my response options: with numbers, verbal labels, or both?	Label all points along your response scale (not just the endpoints) using construct-specific verbal labels.[6,8]	Because of the additional information respondents must process, providing both numbers and verbal labels may increase cognitive effort and can extend response time.
6. Should I state both the positive and negative sides of an issue in the item stem when asking either/or types of questions?	State both sides of an issue in the item stem to reduce bias. For example, in a question about favoring or opposing a new law, it would be better to ask, "Do you favor or oppose the proposed law?" instead of asking, "Do you favor the proposed law?"[8]	Although it is tempting to use fewer words in the item stem by mentioning only one side of an issue, doing so is biased and implicitly suggests there is a more correct answer.

TABLE 2 **Item-Level Statistics for the Pilot Test of the 8 Questions in the Didactic Quality Scale (N = 34 Residents)[a]**

Scale Items	Mean	Mode	Standard Deviation	Range
How do you think the internal medicine integration will impact the educational quality of **morning report**?	3.35	3	0.95	1–5
How do you think the internal medicine integration will impact the educational quality of **noon conferences, including grand rounds**?	3.47	3	0.71	2–5
How do you think internal medicine integration will impact the educational quality of **EKG rounds**?	3.26	3	0.71	2–5
How do you think the internal medicine integration will impact the educational quality of **evidence-based medicine rounds**?	3.32	3	0.68	1–5
How do you think the internal medicine integration will impact the educational quality of **bedside clinical teaching**?	3.26	3	0.57	2–5
How do you think the internal medicine integration will impact the educational quality of **ward team (small-group) didactic experiences**?	3.12	3	0.64	1–5
How do you think the internal medicine integration will impact the educational quality of **procedure-based medical skill proficiency training**?	3.26	3	0.58	2–5
How do you think the internal medicine integration will impact the **availability of educational resources**?	3.26	3	0.75	1–5

Abbreviation: EKG, electrocardiogram.

[a] The 5-point response options were as follows: extreme negative impact, negative impact, neither positive nor negative impact, positive impact, and extreme positive impact.

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

After the draft survey items have undergone expert review, it is important to assess how the target population will interpret the items and response options. One way to do this is through a process known as cognitive interviewing or cognitive pretesting.[12] Cognitive interviewing typically involves a face-to-face interview during which a respondent reads each item aloud and explains his or her thought process in selecting a particular response. This process allows one to verify that each respondent interprets the items as the researcher intended, performs the expected cognitive steps to generate an accurate response, and responds using the appropriate response anchors. Cognitive interviewing is a qualitative method that should be conducted with a handful of participants who are representative of the target population (typically 4–10 participants). This is a critical step to identify problems with question or response wording that may result in misinterpretation or bias. Cognitive interviewing should be conducted using a standardized methodology, and there are several systematic approaches that can be applied.[12]

Given our small target population (approximately 68 residents in all), we opted to perform cognitive interviews with the chief residents and the program and associate program directors of each program, who we felt represented the closest available analogues to our target population. We performed cognitive interviews using both the *think-aloud* and *retrospective verbal probing* techniques.[12] In the think-aloud method, each participant is provided with a copy of the draft survey, which he or she reviews while an interviewer reads from a standardized script. The interviewer reads each item, after which the participant is invited to think aloud while processing the question and selecting a response. Although time consuming, this method of interviewing is helpful in identifying items that fail to evoke the desired cognitive response. Retrospective verbal probing is an alternative method that consists of scripted questions administered just after the participant completes the entire survey. This approach conserves time and allows for a more authentic survey experience; however, retrospective verbal probing can introduce bias related to the participant's memory of each question. Through cognitive interviewing, we identified several small but important issues with our item wording, visual design, and survey layout, all of which were revised in our next iteration.

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

Despite the best efforts of GME researchers during the aforementioned survey design process, some survey items may still be problematic.[4] Thus, to gain additional validity evidence, pilot testing of the survey instrument should be performed. It is important to pilot test the survey using conditions identical or very similar to those planned for the full-scale survey. Descriptive data from pilot testing can then be used to evaluate the response distributions for individual items and scale composite scores. In addition, these data can be used to analyze item and composite score correlations, all of which are evidence of the internal structure of the survey and its relations to other variables. It is also worth noting that other, more advanced statistical techniques, such as factor analysis, can be used to ascertain the internal structure of a survey.[13]

We pilot tested our survey on 14 residents from Walter Reed's IM program and 20 residents from the National Naval Medical Center's IM program; this represented 50% of the target population. TABLE 2 presents the results of our pilot test for the didactic quality scale, which was designed to assess the perceived impact of the IM integration on specific didactic components within each training program. The scale included 8 questions and used a 5-point, Likert-type response scale ranging from "extreme negative impact" to "extreme positive impact."

After reviewing the item-level statistics using SPSS 20.0 (IBM Corp., New York), we calculated a Cronbach alpha coefficient to assess internal consistency reliability of the 8 items in our didactic quality scale. A Cronbach alpha coefficient can range from 0 to 1 and provides an assessment of the extent to which the scale items are related to one another. As we explained in our first editorial, a group of survey items designed to measure a given construct, such as our 8 items designed to measure didactic quality, should all exhibit moderate to strong positive correlations with one another. If they are not positively correlated, this suggests a potential problem with one or more of the items. It should be noted, however, that Cronbach alpha is sensitive to scale length; all other things being equal, a longer scale will generally have a higher Cronbach alpha. As such, a fairly easy way to increase a scale's internal consistency reliability is to add items. However, this increase in Cronbach alpha must be balanced with the potential for more response error due to an overly long survey that is exhausting for respondents.

Although there is no set threshold value for internal consistency reliability, an alpha ≥ 0.75 is generally considered to be acceptable.[14] For our didactic quality scale, the Cronbach alpha was .89, which indicated that our 8 items were highly correlated with one another, as expected. We then calculated a composite score (ie, an unweighted mean score of the 8 items) to create our didactic quality variable, and we inspected the descriptive statistics and histogram of the composite scores (TABLE 2 and FIGURE, respectively).

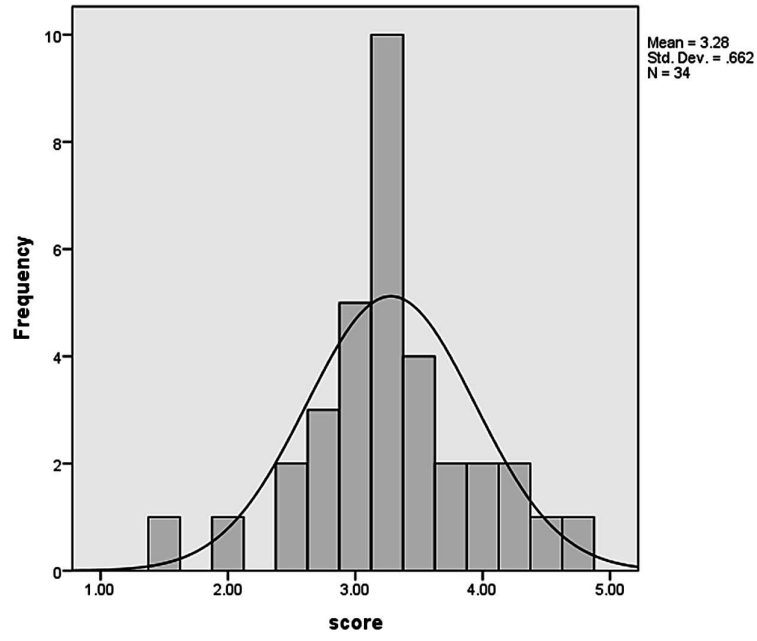

FIGURE **Histogram for the Didactic Quality Composite Variable Scores**

The histogram was normally distributed, which suggested that our respondents were using almost all of the points along our response scale.

After conducting the item- and scale-level analyses, as described above, it is reasonable to advance to the full-scale survey project. Of course, if the pilot results indicate poor reliability and/or surprising relationships (ie, one or more items in a given scale do not correlate with the other items as expected), researchers should consider revising existing items, removing poorly performing items, or drafting new items. If significant modifications are made to the survey, a follow-up pilot test of the revised survey may be in order. If only minor modifications are made, such as removing a handful of poorly performing items, it is reasonable to proceed directly to full-scale survey implementation.

Concluding Thoughts

Developing a high-quality survey takes time. Nevertheless, the benefits of following a rigorous, systematic approach to survey design far outweigh the drawbacks. The example outlined here demonstrates our recommended survey design process. This approach can improve the quality of GME surveys and the likelihood of collecting survey data with evidence of reliability and validity in a given context, with a particular sample, and for a specific purpose. GME researchers are strongly encouraged to follow this or another systematic process when designing surveys and to report validity evidence (ie, the steps of their survey design process) so that readers can critically evaluate the quality of the survey instrument.

References

1 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.
2 Accreditation Council for Graduate Medical Education. ACGME Program Requirements for Graduate Medical Education in Internal Medicine, 2009. https://www.acgme.org/acgmeweb/tabid/134/ProgramandInstitutionalGuidelines/MedicalAccreditation/InternalMedicine.aspx. Accessed November 12, 2012.
3 Bridges W. *Managing Transitions: Making the Most of Change*. 2nd ed. Cambridge, MA: DeCapo Press; 2003.
4 Gehlbach H, Brinkworth ME. Measure twice, cut down error: a process for enhancing the validity of survey scales. *Rev Gen Psychol*. 2011;15:380–387.
5 Gehlbach H, Artino AR Jr, Durning S. AM last page: survey development guidance for medical education researchers. *Acad Med*. 2010;85:925.
6 Artino AR Jr, Gehlbach H, Durning SJ. AM last page: avoiding five common pitfalls of survey design. *Acad Med*. 2011;86:1327.
7 Artino AR Jr, Gehlbach H. AM last page: avoiding four visual-design pitfalls in survey development. *Acad Med*. 2012;87:1452.
8 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.
9 LaRochelle J, Hoellein AR, Dyrbe LN, Artino AR Jr. Survey development: what not to avoid. *Acad Intern Med Insight*. 2011;9:10–12.
10 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(2):94–104.
11 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.
12 Willis GB. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications; 2005.
13 Pett MA, Lackey NR, Sullivan JJ. *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*. Thousand Oaks, CA: Sage Publications; 2003.
14 Gable RK, Wolfe MB. *Instrument Development in the Affective Domain: Measuring Attitudes and Values in Corporate and School Settings*. Boston, MA: Kluwer Academic Publishers; 1993.