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 response options have comparable visual weight, so the 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 nonsubstantive response options together with substantive response options Placing nonsubstantive 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 nonsubstantive response options** Using additional space to visually separate nonsubstantive 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.edu

Table 3. Examples of various Likert-type response options.

Construct being assessed	Five-point, unipolar response scales	Seven-point, bipolar response scales
Confidence	• Not at all confident • Slightly confident • Moderately confident • Quite confident • Extremely confident	• Completely unconfident • Moderately unconfident • Slightly unconfident • Neither confident nor unconfident (or neutral) • Slightly confident • Moderately confident • Completely confident
Interest	• Not at all interested • Slightly interested • Moderately interested • Quite interested • Extremely interested	• Very uninterested • Moderately uninterested • Slightly uninterested • Neither interested nor uninterested (or neutral) • Slightly interested • Moderately interested • Very interested
Effort	• Almost no effort • A little bit of effort • Some effort • Quite a bit of effort • A great deal of effort	
Importance	• Not important • Slightly important • Moderately important • Quite important • Essential	
Satisfaction	• Not at all satisfied • Slightly satisfied • Moderately satisfied • Quite satisfied • Extremely satisfied	• Completely dissatisfied • Moderately dissatisfied • Slightly dissatisfied • Neither satisfied nor dissatisfied (or neutral) • Slightly satisfied • Moderately satisfied • Completely satisfied
Frequency	• Almost never • Once in a while • Sometimes • Often • Almost always	

researchers have the greatest concerns about the scale (relevance, clarity, etc.) for each individual item and for each set of items or scale. The quantitative data combined with qualitative input from experts is designed to improve the content validity of the new questionnaire or survey scale and, ultimately, the overall functioning of the survey instrument.

Step 6: Conduct cognitive interviews

After the experts have helped refine the scale items, it is important to collect evidence of *response process validity* to assess how prospective participants interpret your items and response anchors (AERA, APA & NCME 1999). One means of collecting such evidence is achieved through a process known as cognitive interviewing or cognitive pre-testing (Willis 2005). Similar to how experts are utilized to determine the content validity of a new survey, it is equally important to determine how potential respondents interpret the items and if their interpretation matches what the survey designer has in mind (Willis 2005; Karabenick et al. 2007). Results from cognitive interviews can be helpful in identifying mistakes respondents make in their interpretation of the item or response options (Napoles-Springer et al. 2006; Karabenick et al. 2007). As a qualitative technique, analysis does not rely on statistical tests of numeric data but rather on coding and interpretation of written notes from the interview. Thus, the sample sizes used for cognitive interviewing are normally small and may involve just 10–30 participants (Willis & Artino 2013). For small-scale medical education research projects, as few as five to six participants may suffice, as long as the survey designer is sensitive to the potential for bias in very small samples (Willis & Artino 2013).

Cognitive interviewing employs techniques from psychology and has traditionally assumed that respondents go through a series of cognitive processes when responding to a survey. These steps include *comprehension* of an item stem and answer choices, *retrieval* of appropriate information from long-term memory, *judgment* based on comprehension of the item and their memory and finally *selection* of a response (Tourangeau et al. 2000). Because respondents can have difficulty at any stage, a cognitive interview should be designed and scripted to address any and all of these potential problems. An important first step in the cognitive interview process is to create coding criteria that reflects the survey creator's intended meaning for each item (Karabenick et al. 2007), which can then be used to help interpret the responses gathered during the cognitive interview.

The two major techniques for conducting a cognitive interview are the *think-aloud* technique and *verbal probing*. The think-aloud technique requires respondents to verbalize every thought that they have while answering each item. Here, the interviewer simply supports this activity by encouraging the respondent to keep talking and to record what is said for later analysis (Willis & Artino 2013). This technique can provide valuable information, but it tends to be unnatural and difficult for most respondents, and it can result in reams of free-response data that the survey designer then needs to cull through.

A complementary procedure, verbal probing, is a more active form of data collection where the interviewer administers a series of probe questions designed to elicit specific information (Willis & Artino 2013; see Table 4 for a list of commonly used verbal probes). Verbal probing is classically divided into concurrent and retrospective probing. In concurrent probing, the interviewer asks the respondent specific questions about their thought processes as the respondent answers each question. Although disruptive, concurrent probing has the advantage of allowing participants to respond to questions while their thoughts are recent. Retrospective probing, on the other hand, occurs after the participant has completed the entire survey (or section of the survey) and is generally less disruptive than concurrent probing. The downside of retrospective probing is the risk of recall bias and hindsight effects (Drennan 2003). A modification to the two verbal probing techniques is defined as immediate retrospective probing, which allows the interviewer to find natural break points in the survey. Immediate retrospective probing allows the interviewer to probe the respondent without interrupting between each item (Watt et al. 2008). This approach has the potential benefit of reducing the recall bias and hindsight

Table 1
Checklist of Selected Reporting Guidelines for Studies That Use Surveys

Reporting guideline	Questions to address in the manuscript
Introduction	
1. Provide a rationale for using a survey.[a]	• Why is a survey an appropriate data collection method? • How can the research question(s) be answered using a survey?
Method	
2. Describe how the survey was created or adapted from existing survey(s).	• How were the survey items developed? • What literature was reviewed? • If applicable, what changes were made to previously published surveys?
3. Describe how the survey was pretested prior to full implementation.	• Were experts used to pretest the survey? o If so, describe their qualifications, how many were used, and what the review process was like. • Were cognitive interviews conducted? o If so, describe the interviewees, how many were interviewed, and what the interviewing procedures were like. • Was a pilot test conducted? o If so, describe the sample size, the types of participants, and how the pilot test was conducted.
4. Describe the final survey instrument, including how and when it was administered.	• Has the content of the final survey draft been described in detail (e.g., number and types of items and response options)? • Has a complete, formatted copy of the survey been provided for inclusion in the article's appendix? • What was the method of survey administration (e.g., web or paper based), and where and when was the survey administered? • Were survey responses anonymous or otherwise confidential? • How were respondents contacted, and how often? • How long did respondents have to complete the survey? • Were respondents offered incentives for completing the survey?
Results	
5. Describe the respondents, response rate, and how nonresponse bias was assessed.	• Who comprises the sample, and how does the sample relate to the population of interest? • What was the response rate, and how was it calculated? • Was nonresponse bias assessed, and if so, what was done to correct for it?
6. Describe how score reliability and validity were assessed.[b]	• What processes and statistics were used to assess the reliability of the survey scores? • What sources of validity evidence were collected, and how do they support the intended use of the survey results? o At a minimum, content and response process validity should be considered (e.g., through expert reviews and cognitive interviewing). o So-called "face validity" may be included but should be supplemented by other sources of validity evidence. • If applicable, what type of validity framework was used to guide survey development and validation (e.g., Messick's five sources of validity evidence,[15] Kane's framework[16])?

[a]The rationale can also be presented in the Method.
[b]Reliability and validity evidence are often presented elsewhere in a survey research report (e.g., Introduction, Method, or even Discussion).

Moreover, this rationale should be prominently stated in the manuscript so readers can determine whether a survey is appropriate.

2. Describe how the survey was created or adapted from existing survey(s)

A complete and thorough description of how a survey was created, or how it was adapted from a previously published survey, is a critical component to any survey manuscript. Fortunately, there are guides that describe these processes,[5,9] including how to create good survey items, how to format and administer a survey, and how to analyze the resulting data. In addition, in this issue, Gehlbach and Artino[10] provide a checklist to assist authors in preparing surveys. Evidence-based best practices such as those in the checklist should be consulted and followed; doing so is one of the easiest ways to improve the surveys we use in HPE.

Although it is beyond the scope of this editorial to detail all of the important steps to developing or adapting a survey, we note that authors should be as thorough as possible when describing the processes they used to create their surveys. Important steps in the item-development process include following best practices in item writing and asking content experts to review the items for clarity, relevance, and topic coverage. Even surveys that are adapted from the literature must be closely reviewed by authors and improved upon, where appropriate, since the context for the survey may differ from the circumstances under which it was initially developed. In addition, even instruments pulled from the literature and used "as is" may require additional pretesting prior to use in a new context.

3. Discuss how the survey was pretested prior to full implementation

Some authors do not pretest their surveys prior to use, but they should. Pretesting includes activities like expert reviews, cognitive interviewing, and pilot testing, which can help to establish content and response process validity. Experts can

TABLE
Common Problems and Alternatives in Survey Design

Problem	Poor Example	Better Alternative
1. Leading or biased questions	Which of the following do you believe is *most* responsible for the high costs of health care? ▪ Physicians ▪ Irresponsible health insurance companies ▪ The federal government	Which of the following do you believe is most responsible for the high costs of health care? ▪ Physicians ▪ Health insurance companies ▪ The federal government
2. Double-barreled questions	Was your attending on time to rounds and knowledgeable about the discussion topic?	Was your attending on time to rounds? Was your attending knowledgeable about the discussion topic?
3. Vague questions	How was your rotation experience?	Please rate the quality of attending rounds: Please rate the variety of patients seen on attending rounds:
4. Negatively worded questions	How often do you fail to start rounds on time?	How often do you start rounds on time?
5. Acronyms, nonspecific, or unfamiliar terms	For your most recent MAX experience, were the CBME Milestones reviewed at the start of the rotation?	For your most recent outpatient medicine rotation, were the competencies (skills) for the rotation reviewed at the start of the rotation?
6. Incomplete range or overlapping answer choices	Approximately what percentage of patients who you cared for in clinic the past month were over age 70 years? Choices: 0%–25%; 25%–50%; 50%–75%; 75%–100%	Choices: 0%–24%; 25%–49%; 50%–74%; 75%–100%; not applicable, I was not in clinic last month
7. Absolute answers, such as *always*	In attending rounds over the past 2 months, about how often did you provide scheduled, midrotation formative feedback to the interns? Choices: always; sometimes; rarely; never	Choices: more than 75% of the time; 51% to 75% of the time; about 50% of the time; 25% to 49% of the time; less than 25% of the time
8. Responses that do not match questions	During clinic rotation last month, did the interdisciplinary team meetings assist you in caring for your patients? Choices: strongly agree; agree; neutral; disagree; strongly disagree	Choices: yes, helpful with many patients; yes, helpful with a few patients; no, the team was not helpful; not applicable, I did not discuss my patients with the interdisciplinary team
9. No content review by experts	Survey may omit key areas or not reflect recent studies	Review literature; modify or create the survey; experts review survey
10. No pretesting with similar individuals	Questions and answer responses may not be interpreted consistently by respondents or as you intended	Pretest survey with sample of subjects similar to your target population, using cognitive interviewing techniques[3]
11. No pilot study to examine score reliability or relation to other variables	Survey scores may not be reliable; scores may not measure what you think they measure	Conduct small or large scale pilot study and begin assessing score reliability and validity evidence
12. Excessively long survey	Survey is pages long, with many unnecessary items that may not be used in the analysis	Pretest survey to determine time required to complete; use analysis plan to guide which questions are necessary and which can be removed

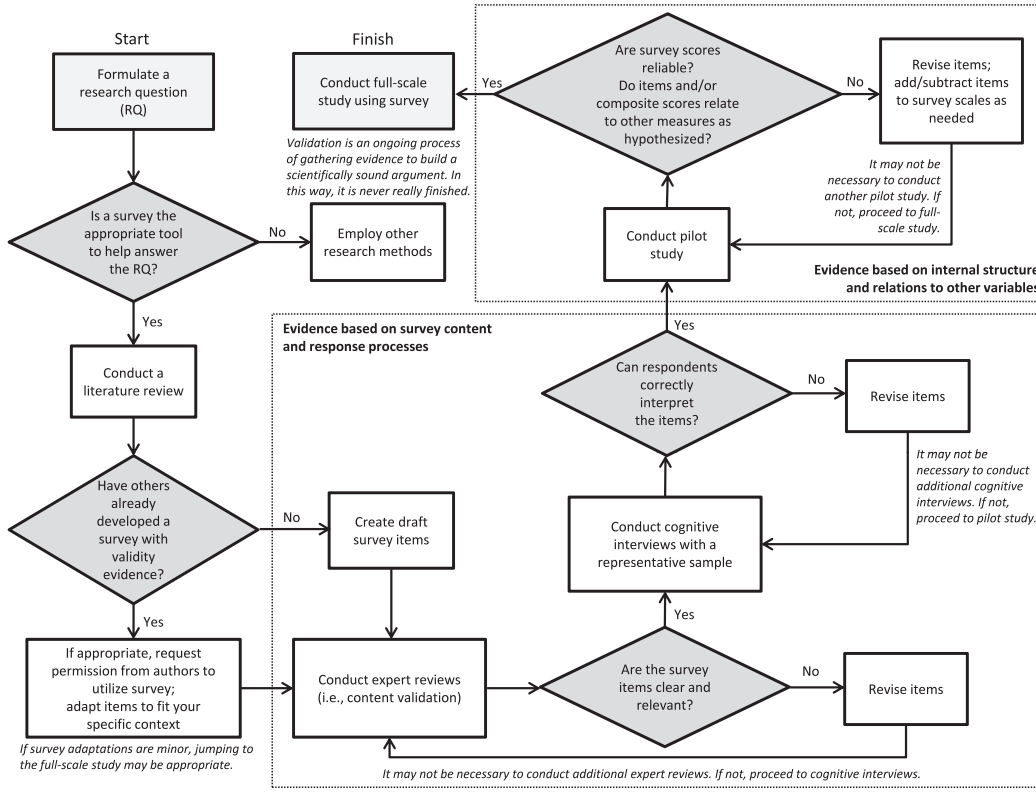

FIGURE **A Systematic Approach to Survey Design for Graduate Medical Education Research**

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