The Survey Checklist (Manifesto)

Hunter Gehlbach, PhD, and Anthony R. Artino Jr, PhD

Abstract

Checklists can mitigate a multitude of high-cost mistakes in fields ranging from surgery to aviation. As part of a standard protocol, checklists may provide many benefits, including improved equity and communication among team members and more efficient integration of different processes during complex tasks. Mostly, though, checklists serve as easy, efficient means to remind professionals of what they already know but can easily forget. By improving processes, checklists can reduce procedural errors, miscommunications, and even deaths. Although the stakes of writing a survey are rarely as high as they are for performing surgery or piloting a plane, checklists can improve the quality of surveys in medical education. In this Perspective, the authors propose a survey checklist to serve the same core function as surgical checklists—to reduce error. That is, a survey checklist can help medical education practitioners and researchers gather more accurate responses. Designers can use the checklist in the appendix to guide item creation processes or to help evaluate the quality of existing surveys. The checklist focuses on formulating items, crafting response options, and formatting/organizing the whole survey.

In his 2010 book, renowned surgeon Atul Gawande[1] illustrates how the wise deployment of checklists (like those used by building contractors and airline pilots) mitigates a multitude of high-cost mistakes in the operating room. As part of a standard protocol, the use of checklists can improve equity during surgeries (e.g., checklists may help nurses feel empowered to remind doctors that they are forgetting a step). Checklists can also facilitate the integration of processes; for example, contractors working on air ducts know to check in with electricians at prescribed moments. Primarily, though, checklists serve as reminders. They help make salient what professionals already know but may forget in moments of crisis. This reinforcement process can reduce procedural errors, miscommunications, infections, and even deaths.[1–3]

Rarely are the stakes of writing a survey as high as they are for performing a surgery. Nonetheless, if medical education practice and research are to improve, it seems essential to collect information from the primary consumers of this education—trainees, practicing physicians, patients, families, and others—in a way that minimizes error and maximizes "true score." Survey error arises from multiple sources—coverage, sampling, nonresponse, and measurement error. Here, we focus on this final category: systematic measurement error, or error arising from poor question wording, formatting, response options, and the order of questions.[4] A recent study found that 95% of medical education surveys violated one or more survey design best practices.[5] Thus, we believe that substantial opportunities to mitigate measurement error exist.

In this Perspective, we present a survey checklist designed to serve the same core function as surgical checklists: to reduce error. Specifically, our hope is that this checklist helps medical education practitioners and researchers reduce measurement error in their survey instruments so that they can gather more accurate answers to their questions of interest. We designed the checklist (see Appendix 1) with an audience of experienced medical education researchers in mind (i.e., those with expertise and training in survey design); however, the checklist may also prove useful in guiding conversations between nonexpert researchers and consultants brought in to help develop survey instruments. Some survey experts may wish to keep the checklist on hand as they develop their own surveys. Others (e.g., those evaluating surveys for possible use) may use the checklist to compare the relative quality (or the relative number of strengths and problems) between two or more survey instruments—possibly even tallying the number of positive/negative responses to produce a heuristic score.

We have restricted the focus of the checklist to issues of formulating items, crafting response options, and formatting/organizing the whole survey. Although other issues such as sampling, response rates, and statistical analysis of survey data are important, they lie beyond our scope. In other words, we presume that a research team using this checklist will have already identified their research question, confirmed that a survey is the right data collection method to fit their needs, and decided on an appropriate corresponding analytic approach. Within this restricted focus, we target issues that are relatively common, cause significant problems, and lack obvious solutions (or the obvious solutions are frequently ignored). Thus, for example, we do not discuss

H. Gehlbach is associate professor, Department of Education, Gevirtz Graduate School of Education, University of California, Santa Barbara, Santa Barbara, California; ORCID: http://orcid.org/0000-0002-2852-2666.

A.R. Artino Jr is professor of medicine and deputy director, Graduate Programs in Health Professions Education, Department of Medicine, F. Edward Hébert School of Medicine, Uniformed Services University of the Health Sciences, Bethesda, Maryland; ORCID: http://orcid.org/0000-0003-2661-7853.

Correspondence should be addressed to Hunter Gehlbach, Department of Education, Gevirtz Graduate School of Education, University of California, Santa Barbara, Santa Barbara, CA 93106-9490; telephone: (805) 893-3385; e-mail: gehlbach@ucsb.edu; Twitter: @HunterGehlbach.

Acad Med. 2018;93:360–366.
First published online November 28, 2017
doi: 10.1097/ACM.0000000000002083

whether or not to include a midpoint by presenting an odd versus even number of response options. Although this issue comes up frequently and lacks an obvious solution (reasonable arguments exist on both sides), the research is equivocal and the consequences of this choice are small.[6]

Formulating Items

When designing survey items, the following best practices typically produce higher-quality data: avoid questions with agree–disagree response items, employ questions with construct-specific response options, ask only one question at a time, use positive language, avoid reverse-scored items, and carefully choose item formats to answer the question asked. These best practices have been derived and tested across more than 40 years of research in the fields of cognitive psychology and public opinion polling.[4,7,8]

Avoid formatting items as statements with agree–disagree response options…

Despite the frequency with which questions with agree–disagree response options are used, survey researchers consistently identify this format as one of the worst ways to present items—for multiple theoretical reasons. This type of question increases both acquiescence bias (the likelihood that respondents will passively agree with whatever statement is presented) and satisficing (i.e., they encourage respondents to put forth suboptimal effort). Further, the precise meaning of the response options tends to be ambiguous. Empirical evidence corroborates that items posed as statements using agree–disagree response options diminish item quality.[9–12]

…And use questions with construct-specific response options instead

Rather than combining statements with agree–disagree response options, the broad consensus among survey researchers is to ask questions and then reinforce the central focus of the question with "construct-specific" response options.[4,12] For example, the question "How much did you enjoy your biochemistry class?" might use the following response options to keep respondents focused on the concept of "enjoyment" as they contemplate their response:

- Did not enjoy at all
- Enjoyed a little bit
- Enjoyed somewhat
- Enjoyed quite a bit
- Enjoyed a tremendous amount

Ask one item at a time (thereby avoiding multibarreled items)

Double- or multibarreled items put respondents in a catch-22 because an individual may endorse one part of a question but reject another.[4,5] For instance, consider the following: "When working with patients, how skilled are you at identifying and accommodating different communication styles?" What happens when a respondent feels skilled at identifying but struggles to accommodate? To avoid trapping respondents in this way, survey designers can ask two separate items—one about identifying and another about accommodating (if both are important). Researchers may also select and ask only the more important question. Another solution is to pose the question at a higher level of abstraction that encapsulates both ideas (e.g., "When working with different types of patients, how skilled are you at adapting your communication style?").

Use positive language

Negative wording—that includes, for example, *un-*, *im-*, *in-*, *anti-*, or *not*—proves challenging for respondents in two ways. First, because negatives are often signaled by just a few letters, they can be easily overlooked by respondents who are rushing through a survey. Second, even if respondents do read them, conjuring up the mental image of the absence or opposite of something is cognitively challenging. Wegner's work[13] on ironic processes illustrates this challenge: His research shows that telling people, "Don't think about pink elephants" is especially likely to cause them to conjure up images of pink elephants. Thus, negatively worded survey items often cause respondents to make easily avoidable errors in reading, considering, and answering items. For instance, an item such as "How inappropriate is it for attending physicians to use the technique of resident pimping during bedside rounds?" might be reworded to "How appropriate is it for attending physicians to use the technique of resident pimping during bedside rounds?"

Avoid "reverse-scored" items

In a related practice, some survey designers advocate the use of reverse-scored items as a means of keeping survey takers alert. The idea is to insert items into a scale whose valence is the opposite of the other items on the scale so that respondents must read each item carefully. For example, imagine the item "How often were you annoyed during your clinic visit?" within a visit satisfaction scale in which the other items inquire about positive elements of the visit (friendliness, punctuality, and other qualities where higher scores would correspond to greater satisfaction). In theory, these reverse-scored items should help respondents realize that they need to focus carefully; they cannot slip into autopilot mode. However, in practice, these items typically diminish the reliability of the overall scale[14]—especially when used with less-educated respondents (such as children[15]).

Choose item formats that answer the questions of interest

Although selecting item formats that actually answer the questions asked seems painfully obvious, survey designers far too frequently structure items in ways that fail to yield the data they need. Sometimes they choose an intrinsically faulty item structure. For example, check-all-that-apply items routinely result in respondents picking more items toward the top of the list and disproportionately ignoring later items. As a result, whether particular items do not apply or whether an inattentive respondent skipped them is unclear.[10] A better approach is a forced-choice format, which encourages respondents to provide more complete, high-quality data. See Box 1 for an illustration. Using this type of forced-choice format, researchers will always know whether a respondent answered or skipped a particular item.

At other times, researchers might present respondents with rating items when their research question clearly calls for rank-ordered data. For example, when asking faculty members to prioritize a list of five highly desirable university initiatives, a ranking question may be more useful to the researcher because it allows respondents to identify which of the initiatives are most and least preferred.

Sometimes, researchers may realize that open-ended/free-response items better

Box 1

Illustration of a "Forced-Choice" Item Format

Survey respondents routinely select more of the responses towards the top of the list in check-all-that-apply formats and disproportionately ignore later choices. A better format is forced-choice, which encourages respondents to provide a response for each item. The approach shown here informs researchers whether respondents answered or skipped an item. For example: "Which of the following medical education topics are you interested in studying during your degree program? Please check *Yes* or *No* for each one."

Yes	No	
○	○	Teaching and learning
○	○	Curriculum development
○	○	Evaluation and assessment
○	○	Research methods
○	○	Leadership and management

match their research questions than closed-ended questions. The key point is to align item formats with the research question being asked so that the right data are obtained.

Crafting Response Options

The following reminders on crafting response options can further help mitigate respondent error: choose an appropriate number of response options; label all response options; use only verbal labels; balance the visual, numeric, and conceptual midpoint of the response options; visually separate nonsubstantive choices from other response options; and format response options into only one row or one column. Note that the first and final guidelines refer to both rating and ranking items; the remaining guidelines typically apply to rating items only.

Choose an appropriate number of response options

For rating items, the number of response options requires striking a balance between providing enough options for participants to precisely represent their opinions, attitudes, or behaviors and providing few enough options for participants to clearly distinguish the meaning of each. The right balance will depend on the topic, the knowledge and cognitive sophistication of the respondents, and the analyses planned. That said, some scholars have proposed that the optimal number of response options typically lies between four and seven. Krosnick and Fabrigar[16] suggest that the optimal number of options may be five for unipolar continua (i.e., continua that array from a conceptual zero point to infinity, such as the frequency of a behavior) and seven for bipolar continua (i.e., continua that array from negative to positive infinity, such as positive or negative attitudes). Relatedly, Weng's findings[17] suggest that only four response options may be too few (at least for college-aged respondents).

For ranking items, the appropriate number of response options is more a function of the cognitive sophistication of respondents. For example, well-educated adults may be able to rank six different responses (e.g., about their dietary preferences), but this task may exceed the capacities of many pediatric patients. For some research questions, asking respondents to rank only a portion of the response options (e.g., please rank the top three) may be a reasonable alternative.[4,18]

Label all response options

By ensuring that each response option for a rating item has a verbal label (as opposed to leaving some options blank), survey designers not only help each option seem equally viable but also ensure that the meaning of each option is clear.[19] Labeling all response options will improve the odds that all respondents will interpret each response option in the same way. See Box 2 for a contrast between fully and partially labeled response options.

Use only verbal labels

Intuition might dictate that numbers provide greater precision than words; however, studies have shown that in self-administered surveys, verbal labels hold more consistent meanings from person to person than numbers.[19] In other words, one person's "4" may be quite different than another's, but what one person means by "quite painful" is typically close to what someone else means. Thus, for response options for rating items, words provide more clarity than numbers.

Balance the visual, numeric, and conceptual midpoint of the response options

How survey designers visually array response options influences which option people choose. Survey designers may confuse respondents if the visual balance of the response scale, number of response options, and conceptual meaning of the options are not completely congruent.[20] For example, the first block of response options in Box 3 allows respondents to reasonably conclude any of the following:

- The line between "good" and "very good" is the midpoint (*visually* this is true because the fourth and fifth response options take up so much room),
- "good" is the midpoint (*numerically* it is the third option on a five-point scale), or

Box 2

Illustration of the Contrast Between Item Formats That Do and Do Not Provide Fully Labeled Response Options

By providing a verbal label for each response option (as opposed to leaving some options blank), survey designers help each option seem equally viable and clarify the meaning of each option. To illustrate, the item asking: "How skilled are you at suturing?" will work better followed by these response options:

not at all skilled	slightly skilled	moderately skilled	quite skilled	extremely skilled

…as opposed to these response options:

not at all skilled				extremely skilled

Box 3

Illustration of Formats That Do Not and Do Balance the Visual, Numeric, and Conceptual Midpoint of the Response Options

Some survey designers inadvertently confuse respondents when the visual balance of the response scale, number of response options, and the conceptual meaning of the options are not completely congruent. For example, the scale shown here:

poor	fair	good	very good	excellent

allows respondents to reasonably conclude any of the following:

- the line between "good" and "very good" is the midpoint (*visually* the three responses to the left of this line consume the same amount of space as the two responses to the right of this line),
- "good" is the midpoint (*numerically* it is the third option on a five-point scale), or
- "fair" is the midpoint (*conceptually* "fair" connotes neither good nor bad).

Instead, survey designers should strive to ensure that the visual, numeric, and conceptual midpoints of a set of response options all align as in the example shown here:

very negative	moderately negative	slightly negative	neither negative nor positive	slightly positive	moderately positive	very positive

- "fair" is the midpoint (*conceptually* "fair" connotes neither good nor bad).

The midpoint strongly signals what each of the other response options means, and confusion as to where the midpoint actually lies can introduce substantial error because different respondents will rely on different interpretations. The key, then, is to ensure that the visual, numeric, and conceptual midpoint of a set of response options all align, as in the second block of response options in Box 3. In this case, the visual midpoint of the response options lies directly over the middle of the "neither/nor" response option. What's more, this point is numerically the fourth option on the seven-point response scale and is conceptually equidistant between very negative and very positive.

Visually separate "nonsubstantive" choices from the other response options

An important exception to the guideline that response options be evenly spaced is when one or more response options (e.g., "don't know" or "not applicable") lack substantive meaning. By visually distinguishing these options from the main substantive responses with extra space, respondents can clearly infer which responses are part of the underlying continuum on a rating scale and where the midpoint of that continuum lies.[20] See the illustration in Box 4.

Format your response options into only one row or only one column

Whether asking respondents to rate or rank response options, forcing them to read from both top to bottom and left to right increases confusion and causes more error. Instead, survey designers should use a single row or a single column to array the response options.[21] Although no extant research appears to suggest a clear differential benefit of either rows or columns, columns may be preferable as respondents increasingly take surveys on their smartphones and smartphone screens are, by default, vertically oriented.[22]

Box 4

Illustration of Formats That Do Not and Do Visually Separate "Nonsubstantive" Choices From the Other Response Options

To preserve the clarity of what the midpoint of a block of response options is, survey designers will want to visually separate the "non-substantive" response options from the main continuum of substantive response options. For example, instead of this approach:

not at all exciting	mildly exciting	somewhat exciting	quite exciting	extremely exciting	not sure

A better approach is:

not at all exciting	mildly exciting	somewhat exciting	quite exciting	extremely exciting		not sure

Putting It All Together: Formatting and Organizing the Whole Survey

The following practices regarding the formatting and organization of items within a larger survey will help maintain respondents' engagement and effort and, in turn, will help generate high-quality data: Ask more important questions earlier in the survey, ensure that each item applies to every respondent, use scales rather than single items when possible, format the visual layout of the instrument consistently, and place sensitive items toward the end of the survey.

Ask the more important items earlier in the survey

Placing the most important items at the beginning of the survey increases the odds that respondents will answer these items while they have ample energy and focus.[4] For example, an instrument about medical student satisfaction with a course should ask about the instructor and the learning activities early in the survey, leaving less-relevant questions about a respondent's prior education for later. Moreover, the first question on a survey is arguably the most important, especially on Internet-based surveys, because it often determines whether or not invited participants complete the survey.[4] A good first question is fairly simple; applies to all respondents; is easy to read, comprehend, and answer; and aligns with the purpose of the survey as described in the invitation.[4]

Ensure that each item applies to every respondent

Establishing and maintaining rapport with respondents is immensely important. Without a strong, trusting relationship, respondents will put forth less effort and likely provide low-quality answers (if they respond at all). One way a survey designer

may inadvertently alienate respondents is by asking numerous items that do not apply to them. The cost of asking low-relevance items tends to be higher at the beginning of surveys where respondents may be more likely to quit the survey altogether. Using branching items that route respondents to only relevant items is typically a better strategy.[4] Although branching items present modest logistical challenges for paper-and-pencil surveys, they can be easily created through most web-based survey programs (e.g., Qualtrics or SurveyMonkey) to provide a seamless user experience. For example, rather than asking, "How satisfied were you with the resources available in the medical library" and giving respondents a "not applicable" option, ask them first, "Have you used the medical library's resources?" If they respond "yes," then ask them how satisfied they were with those resources.

Use scales rather than single items when possible (especially for more complex topics)

A scale consists of several related survey items that, as a group, are designed to measure the same underlying idea or construct.[23–25] Although scales take longer for respondents to complete, they bolster accuracy (compared with error-prone single items), particularly when assessing complex topics. For example, if assessing a patient's overall clinic experience, the question "How friendly was your doctor during your clinic experience?" measures only a portion of the construct of interest. On the other hand, a five-item satisfaction scale that assesses a representative cross-section of the experience should improve measurement. To illustrate, asking all of the following items will provide more precise data:

1. How welcoming were the front desk personnel?
2. During your visit, how satisfied were you with the cleanliness of the clinic?
3. How helpful was the interaction you had with your physician?
4. In general, how efficient was the entire clinic visit?
5. Overall, how satisfied were you with your clinical care?

Properly developing scales to ensure not only that the set of items constituting a scale adequately represents an underlying construct but also that construct-irrelevant questions are not included is a prerequisite for reducing measurement error relative to single items.[26]

Ensure that the visual layout of the survey is consistent

A consistent visual layout teaches respondents where to look for vital information on a survey in a quick, efficient fashion.[4] In essence, consistency teaches them the most proficient way to process the survey. Changes in visual layout or instructions force respondents to take time to relearn what is being asked and how they should answer. Commonly, survey designers find themselves changing the visual layout of their survey in an effort to cram as much content as possible onto a single page. Although keeping a survey short is a laudable goal, on balance, using a consistent visual layout—and, if necessary, more pages—is better than sacrificing clarity by formatting a survey like a jigsaw puzzle.

Place sensitive items, such as demographic questions, later in the survey

Many respondents feel uncomfortable divulging demographic information. Furthermore, some respondents might respond differently on subsequent items if their race, gender, or social class is made salient, as a by-product of stereotype threat.[27] Consequently, the consistent recommendation from survey design experts is to ask for demographic data (or other sensitive information) toward the end of surveys.[4] Collecting personal data later helps ensure that respondents do not alter their answers as a result of particular sensitive items being asked first. Furthermore, this strategy allows most other data to be collected—even if a respondent takes offense at the sensitive items and quits the survey early. For example, residents completing a graduate medical education survey might abandon the survey before they even begin if the first item asks, "How often did you break the rule that there should be a 10-hour time period between all daily duty periods and after in-house call?"[28] The question is clearly sensitive and should not be asked until more rapport and trust have been built as respondents become more interested and engaged with the survey.[6]

Limitations and Other Considerations

Although we are confident that addressing the items on our checklist will help mitigate substantial sources of measurement error in medical education surveys, they are far from a survey researcher's only considerations. For instance, researchers selecting an "off-the-shelf" survey will want to weigh the evidence of validity that exists for each potential scale within the survey in light of their target population, context, intended application, and so forth.[29] For researchers developing new surveys, this checklist addresses best practices surrounding only the writing of items. The larger survey design process, the importance of pilot testing (including techniques such as cognitive interviewing and expert reviews), and approaches to creating composites and analytic techniques all go beyond the scope of this checklist.[26,30,31]

Furthermore, we have limited the checklist only to topics for which reasonably robust evidence is available. Recommendations regarding how many questions per page work best or how long to make a survey would be valuable. However, as experts such as Krosnick[32] have identified, these and a vast number of other considerations (e.g., the cognitive sophistication and stamina of respondents, online vs. mobile phone vs. paper/pencil modes of administration, strength of Internet connection, and respondent motivation) all require further investigation before researchers can arrive at reasonable prescriptions. Likewise, context effects—that is, the way certain survey items influence respondents' answers to surrounding items—have become such a notable topic within survey research[4,8,33] that some scholars have turned them into interventions.[34] However, because they are so idiosyncratic and depend so much on the specific context of the survey, avoiding these effects remains more art than science at present.

In Sum

A checklist serves as a simple way to remind professionals of what they already know but can easily forget. Studies suggest that surgical checklists can improve communication[2] and reduce morbidity and mortality.[3] Although the stakes for writing surveys are rarely as high as those for performing surgery, we believe the smart use of a checklist such as this one can reduce measurement error

in medical educators' surveys—just as checklists can reduce error in surgeons' performance.

Funding/Support: None reported.

Other disclosures: Dr. Hunter Gehlbach serves as the director of research at Panorama Education. He has previously advocated and articulated the approach presented in this Perspective. Panorama is a company that helps K–12 educators and school districts develop better surveys.

Ethical approval: Reported as not applicable.

Disclaimer: The views expressed in this Perspective are those of the authors and do not necessarily reflect the official policy or position of the Uniformed Services University of the Health Sciences, the U.S. Navy, the Department of Defense, or the U.S. government.

Previous presentations: The checklist presented in this Perspective is an adaptation of an earlier tool developed by Dr. Gehlbach. It is available at Panorama Education: https://panorama-www.s3.amazonaws.com/files/survey-resources/checklist.pdf. Although the checklist presented here is similar to the tool published on the Panorama Education website, the content of the Perspective itself is completely new.

References

1 Gawande A. The Checklist Manifesto: How to Get Things Right. New York, NY: Metropolitan; 2010.
2 Lingard L, Regehr G, Orser B, et al. Evaluation of a preoperative checklist and team briefing among surgeons, nurses, and anesthesiologists to reduce failures in communication. Arch Surg. 2008;143:12–17.
3 Haynes AB, Weiser TG, Berry WR, et al.; Safe Surgery Saves Lives Study Group. A surgical safety checklist to reduce morbidity and mortality in a global population. N Engl J Med. 2009;360:491–499.
4 Dillman DA, Smyth JD, Christian LM. Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method. 4th ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2014.
5 Artino AR, Phillips AW, Utrankar A, Ta AQ, Durning SJ. "The questions shape the answers": Assessing the quality of published survey instruments in health professions education research. Acad Med. 2018;93:456–463.
6 Nadler JT, Weston R, Voyles EC. Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. J Gen Psychol. 2015;142:71–89.
7 Krosnick JA, Presser S. Question and questionnaire design. In: Marsden PV, Wright JD, eds. Handbook of Survey Research. 2nd ed. Bingley, UK: Emerald Group Publishing; 2010.
8 Schwarz N. Self-reports: How the questions shape the answers. Am Psychol. 1999;54:93–105.
9 Fowler FJ Jr. Survey Research Methods. 5th ed. Thousand Oaks, CA: Sage Publishing; 2014.
10 Gehlbach H. Seven survey sins. J Early Adolesc. 2015;35:883–897.
11 McIntrye J, Gehlbach H. The cost of agree–disagree: Satisficing and sacrificing reliability. Paper presented at: Society of Research on Educational Effectiveness; March 7, 2014; Washington, DC. http://files.eric.ed.gov/fulltext/ED562897.pdf. Accessed November 9, 2017.
12 Saris WE, Revilla M, Krosnick JA, Shaeffer EM, Shaeffer EM. Comparing questions with agree/disagree response options to questions with item-specific response options. Surv Res Methods. 2010;4:61–79.
13 Wegner DM. Ironic processes of mental control. Psychol Rev. 1994;101:34–52.
14 Swain SD, Weathers D, Niedrich RW. Assessing three sources of misresponse to reversed Likert items. J Mark Res. 2008;45:116–131.
15 Benson J, Hocevar D. The impact of item phrasing on the validity of attitude scales for elementary school children. J Educ Meas. 1985;22:231–240.
16 Krosnick JA, Fabrigar LR. Designing rating scales for effective measurement in surveys. In: Lyberg L, Biemer P, Collins M, et al, eds. Survey Measurement and Process Quality. New York, NY: John Wiley & Sons, Inc.; 1997:141–164.
17 Weng L-J. Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. Educ Psychol Meas. 2004;64:956–972.
18 Krosnick JA. Maximizing questionaire quality. In: Robinson JP, Shaver PR, Wrightsman LS, eds. Measures of Political Attitudes. San Diego, CA: Academic; 1999:37–57.
19 Krosnick JA. Survey research. Annu Rev Psychol. 1999;50:537–567.
20 Tourangeau R, Couper MP, Conrad F. Spacing, position, and order: Interpretive heuristics for visual features of survey questions. Public Opin Q. 2004;68:368–393.
21 Christian LM, Parsons NL, Dillman DA. Designing scalar questions for Web surveys. Sociol Methods Res. 2009;37:393–425.
22 De Bruijne M, Wijnant A. Improving response rates and questionnaire design for mobile web surveys. Public Opin Q. 2014;78:951–962.
23 DeVellis RF. Scale Development: Theory and Applications. 3rd ed. Thousand Oaks, CA: Sage; 2012.
24 Gehlbach H, Artino AR Jr, Durning S. AM last page: Survey development guidance for medical education researchers. Acad Med. 2010;85:925.
25 Artino AR Jr, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE guide no. 87. Med Teach. 2014;36:463–474.
26 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.
27 Steele CM, Aronson J. Stereotype threat and the intellectual test performance of African Americans. J Pers Soc Psychol. 1995;69:797–811.
28 Accreditation Council for Graduate Medical Education. Resident/fellow and faculty surveys. http://www.acgme.org/Data-Collection-Systems/Resident-Fellow-and-Faculty-Surveys. Updated 2017. Accessed November 11, 2017.
29 Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50:741–749.
30 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.
31 Willis GB, Artino AR Jr. What do our respondents think we're asking? Using cognitive interviewing to improve medical education surveys. J Grad Med Educ. 2013;5:353–356.
32 Krosnick JA. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cogn Psychol. 1991;5:213–236.
33 Gehlbach H, Barge S. Anchoring and adjusting in questionnaire responses. Basic Appl Soc Psych. 2012;34:417–433.
34 Gehlbach H, Robinson CD, Finefter-Rosenbluh I, Benshoof C, Schneider J. Questionnaires as interventions: Can taking a survey increase teachers' openness to student feedback surveys? [published online ahead of print July 25, 2017]. Educ Psychol. doi: 10.1080/01443410.2017.1349876.

Appendix 1

Survey Design Checklist

For formulating items: *Does your survey…*

	Yes (1 point)	No (0 points)
Avoid formatting items as statements with agree/disagree response options…		
…And use questions with construct-specific response options instead		
Ask one item at a time (thereby avoiding multibarreled items)		
Use positive language (i.e., avoid *un-*, *in-*, *anti-*, etc.) to ease cognitive processing		
Avoid "reverse-scored" items		
Use item formats that answer your research questions of interest		
Formulating items subscore =	**/ 6**	

For crafting response options: *Does your survey…*

	Yes (1 point)	No (0 points)
Use an appropriate number of response options		
Include labels for all response options		
Use only verbal labels		
Balance the visual, numeric, and conceptual midpoint of the response options		
Visually separate nonsubstantive choices from the other response options		
Provide response options in only one row or only one column		
Response options subscore =	**/ 6**	

For formatting and organizing the whole survey: *Does your survey…*

	Yes (1 point)	No (0 points)
Ask the more important items earlier		
Include only items that apply to every respondent (or employ branching items)		
Use scales—not single items—when possible (especially for complex topics)		
Use a consistent visual layout		
Place sensitive items, such as demographic questions, later		
Formatting/organizing subscore =	**/ 5**	

Total score[a] =	**/ 17**

[a]Although there is no absolute target score, a quick tally of the number of "yes" check marks should allow for the comparison of different survey instruments.