Review of General Psychology
2011, Vol. 15, No. 4, 380–387

1089-2680/11/$12.00 DOI: 10.1037/a0025704

Measure Twice, Cut Down Error: A Process for Enhancing the Validity of Survey Scales

Hunter Gehlbach and Maureen E. Brinkworth
Harvard Graduate School of Education

For years psychologists across many subfields have undertaken the formidable challenge of designing survey scales to assess attitudes, opinions, and behaviors. Correspondingly, scholars have written much to guide researchers in this undertaking. Yet, many new scales violate established best practices in survey design, suggesting the need for a new approach to designing surveys. This article presents 6 steps to facilitate the construction of questionnaire scales. Unlike previous processes, this one front loads input from other academics and potential respondents in the item-development and revision phase with the goal of achieving credibility across both populations. Specifically, the article describes how (a) a literature review and (b) focus group–interview data can be (c) synthesized into a comprehensive list to facilitate (d) the development of items. Next, survey designers can subject the items to (e) an expert review and (f) cognitive pretesting before executing a pilot test.

Keywords: survey research, research methodology, assessment, measurements, validity–reliability, scales

The proliferation of new questionnaire scales across different domains of psychology has continued for decades (Clark & Watson, 1995; Comrey, 1988; Simms & Watson, 2007). Concurrently, the scholarship advising researchers on the best ways to construct questionnaires has also grown over time. Yet, illustrations abound of new scales that deviate from these best practices in questionnaire design. Why this paradox exists is not entirely clear. Perhaps a paucity of graduate training on survey design is to blame; perhaps too much of the scholarly guidance on questionnaire development remains sequestered in methodology journals thereby eluding those who actually design surveys; or perhaps long-standing habits of developing scales through the solitary Likert approach (McIver & Carmines, 1981) die hard, despite the availability of new techniques. Presumably, some truth resides in each of these explanations.

This article introduces six steps to designing surveys that we hope will improve the development of survey scales within psychology and other social sciences regardless of which combination of these reasons is most accurate. Although several ideas in this article are new, our primary goal is not to present new techniques in survey design. Rather we strive to synthesize several known (though not necessarily widely known) survey design practices to create a new process that differs importantly from previous processes (e.g., Clark & Watson, 1995; Comrey, 1988; Simms, 2008) in three main ways. First, this process balances the use of diverse techniques rather than relying heavily on psychometric analyses. Second, this process is inherently collaborative and relies on other experts in the field as well as potential participants. Finally, this process front loads the task of establishing validity. By focusing on validity during the development of items, there is high potential for more efficient (i.e., shorter) scales as well as more efficient pilot testing. This process may not fully endow survey designers with the economy of carpenters who can "measure twice and cut once," but it is likely to reduce measurement error and enhance the validity of most new scales.

Although many of the techniques we describe apply to the development of individual survey items, we focus on scale development—that is, a series of similar items designed to assess the same underlying construct that are then summed to represent a participant's score on the construct (DeVellis, 2003). Scales are more complicated to develop and take more respondent time. However, they pay dividends to researchers by more fully, precisely, and reliably assessing the underlying construct (McIver & Carmines, 1981). As a consequence of these virtues, scales are widely used across most subfields within psychology.

We begin by describing common scale construction processes to provide a backdrop to the process we advocate. Specifically, we highlight the strengths in the typical process and describe how complementing these techniques with more preliminary work in the item-development phase should enhance the validity of the final survey scales. Then we describe six steps to facilitate the design of survey scales. To illustrate the value of each step, we pull examples and illustrations from parallel scales that we constructed recently to assess teacher–student relationships (TSR) from both teachers' and students' perspectives at the middle- and high-school levels.

This article was published Online First November 7, 2011.

Hunter Gehlbach and Maureen E. Brinkworth, Harvard Graduate School of Education, Harvard University.

This research was made possible in part through generous support from the Harvard Graduate School of Education Dean's Venture Fund. Many of these techniques have been progressively refined through Harvard Graduate School of Education's course on survey design—those students are thanked for their patience and support through all of the tinkering. Thanks also to Hahrie Han and Jessica Scott for their feedback on drafts of this article.

Correspondence concerning this article should be addressed to Hunter Gehlbach, Harvard Graduate School of Education, Longfellow Hall 328, 13 Appian Way, Cambridge, MA 02138. E-mail: Hunter_Gehlbach@gse.harvard.edu

A "Typical" Survey Design Process

Several authors have provided excellent guidance for scholars who are interested in developing surveys (e.g., Clark & Watson, 1995; DeVellis, 2003; Simms, 2008). Though each recommended process is unique, several commonalities extend through these and other approaches. First, the goal of these processes is generally the same—to produce a scale that demonstrates evidence of construct validity. Construct validity comes in many forms—content, substantive, structural, generalizability, external, and consequential—and is a function of the items within a survey scale, the population of respondents, the context in which they are taking the survey, and the ultimate use of the scores derived from the survey (Messick, 1995). Thus, validity is not so much an endstate (i.e., the idea of a "validated scale" is a misnomer) as a property of scales, which scholars can find increasing amounts of evidence for (or against) for different respondents in different contexts and for different uses of the resultant scores. Among those who have proposed processes for scale development, there is broad consensus that for a scale to adequately measure the construct that it purports to measure, it must minimize respondent error and attain a certain level of reliability. Our process for scale construction shares this basic goal of developing a scale with substantial evidence of construct validity.

Second, looking across these processes for designing surveys, many recommendations are similar. The general template suggests that survey designers should (a) clearly determine the construct in question, (b) consult the literature to ascertain whether a new scale is needed, (c) develop an item pool (that should be overly inclusive), (d) select appropriate response formats for items, and (e) conduct several iterations of pilot testing. The pilot testing should include the focal items as well as items that begin to establish validity for the scale and be followed by analyses to eliminate problematic items. The bulk of these processes center on the analytic approaches a researcher can take after having collected pilot data from a preliminary sample. This psychometric toolkit includes analyses to assess: item-level means and variability; interitem and item-total correlations; reliability (e.g., coefficient alpha or test–retest); factor structure (e.g., exploratory and/or confirmatory factor analysis); multitrait, multimethod matrix approaches to establishing validity; item-response theory; and so on. In sum, these processes emphasize the back-end of scale development—that is, the selection of items—while focusing less on the development of items.

Each of these steps and analytic techniques are tremendously important to the development of high-quality scales, and we rely heavily on each of them in our own work. At the same time, we argue that this "typical" survey design process could be improved. Specifically, a process that (a) uses a broader range of techniques, (b) encourages scholars to be more collaborative with other researchers and with potential respondents during item development, and (c) increases the emphasis on validity early in the process should ultimately produce more efficient, valid scales while requiring fewer pilot tests.

Step 1: Literature Review

Congruent with most scale construction processes, our scale construction process begins with a thorough literature review. However, this step is more than a mere search for an extant measure that might serve to assess the construct in question. This step has two goals: to precisely define the construct in relation to literature and to identify how existing measures of the construct (or related constructs) might be useful. First, knowledge of the literature helps survey designers define their construct so as to situate it within, connect it to, and differentiate it from related concepts. A new practice that some researchers may find helpful is to sketch Venn diagrams that illustrate the degree of overlap between their construct of interest and related, but distinct, constructs. These diagrams help illustrate what the construct is (or is not) and provide a later reference check for evaluating the construct relevance of items. Of particular importance, these diagrams can help scholars determine the appropriate "grain size" of their construct, that is, the level of abstraction at which to measure their construct. For example, to investigate the social proclivities of undergraduates, one could develop scales to assess their propensity to go on dates (at the small-grain end of the continuum), spend time with friends, or interact with others (at the large-grain end of the continuum). Depending on whether one wanted to predict sexual behavior, use of Facebook, or rates of depression, scales of different grain sizes would serve as more or less effective predictors.

In our examination of the TSR literature we observed that past research often examined small-grain subcomponents of TSR (e.g., Wentzel, 1997, and her work on teacher caring), assessed the teachers' (e.g., Birch & Ladd, 1998) or students' (e.g., Goodenow, 1993) perceptions of the relationships (though rarely both), and more frequently focused on elementary school students (Pianta, 1999). Thus, despite the excellent work in this field, we saw ways in which a more comprehensive (i.e., large-grain) measure of TSR that captured both teacher and student perceptions at the secondary level could make important scientific and practical contributions.

Our final conceptualization of TSR—that they are a dynamic and reciprocal social process consisting of students' and teachers' interpersonal interactions and their respective perceptions of those interactions—drew heavily from Pianta's (1999) work. However, our large-grain approach and need for parallel teacher and student measures distinguished our measure from previous ones. The Venn diagram technique described above helped us sharpen the precision of our construct. For example, our diagram helped us clarify that one party's liking of the personality of the other party was part of our construct, however, the specific personality traits of each party were not.

Once researchers clarify their construct of interest, reviewing the literature also serves to evaluate previous measures for potential use. Scale validation studies, methods sections, and appendixes typically provide sample items, the full scale, or both, as well as psychometric properties of the scale. In some instances, scholars may find scales that closely match their construct of interest and can use them with only minor modifications. In many cases, the items themselves might be unusable (e.g., the reading level of the items is inappropriate or the items do not comport with best practices). However, the content these items address might be valuable in developing new items. As scholars become familiar with the best practices in designing items described in Step 4, they will become proficient at adjudicating the extent to which scales might be borrowed with minor revisions versus requiring substantive changes.

As we combed through prior measures of TSR, some items required only slight modifications. For example, one item, "I like this student," from the Teacher-Student Relationship Inventory (Ang, 2005) was reworded to "How much do you like Student X's personality?" In other instances, certain scales contained content that was critical to our conceptualization of TSR, but we changed the wording substantially. For instance, "I share an affectionate, warm relationship with this child" from the Student-Teacher Relationships Scale (STRS; Pianta, 2001) addresses closeness, an important subcomponent of TSR. As it is worded, however, this item seemed inappropriate for use with older students, but we retained this general concept for the next step of the process.

Step 2: Interviews and Focus Groups

With the literature review completed, researchers can turn their attention to the population of interest—an important deviation from most traditional survey construction processes. Specifically, scholars need to ascertain whether their newly refined conceptualization of the construct matches the way their prospective respondents think about it. Do respondents include and exclude the same categories as those in the literature? What terminology do respondents use in describing relevant phenomena? To answer these questions, researchers will usually want to collect data directly from individuals who closely resemble their population of interest. Scholars may use one-on-one interviews or focus groups (or a combination) toward this end. Regardless of the approach taken, these discussions should be structured around two main objectives. First, researchers need to hear how participants think about the focal construct in their own words, with minimal prompting from the researcher. Second, after getting as much unprompted information as possible, survey designers can ask more directed, probing questions to assess whether respondents agree with certain characteristics of the construct noted in the literature. Scholars will want to continue until additional interviews–focus groups begin yielding little new information in regard to potential respondents' conceptualization of the construct.

At Step 2 in our investigation of TSR, we interviewed teachers and conducted focus groups with students. For the latter group, we first asked them, "I'd like you to think about what it means to have a good/positive relationship with a teacher. Now can you describe what a good relationship with a teacher is to you—what does it look like?" Next, we asked participants to talk about a teacher with whom they had a particularly good relationship; especially for the younger students, we felt it was important to make the task as concrete as possible. Once respondents had exhausted most of what they could tell us unprompted, we shifted into more directive questions such as, "Was there something about how the teacher taught the class—did s/he allow you to do more independent or self-directed work?" (an attempt to learn about the role of student autonomy). Our approach with teachers followed the same trajectory of transitioning from very open-ended questions to more focused prompts.

Because our construct was multifaceted, we devised an adaptation of the Q-sort technique to efficiently learn which categories from our literature review students and teachers felt were important. Toward the end of the interviews, participants completed a Q-sort procedure in which they sorted cards labeled with aspects of TSR that we had culled from our literature review. Specifically, they responded to the question, "How important are each of these characteristics to developing a positive relationship with your teacher/students?" To respond, participants grouped cards into the following categories: *extremely important/essential*, *somewhat important*, and *not important/doesn't matter*. One interesting result from this procedure occurred around the idea of responsiveness. According to the literature, it is critical for teachers to be responsive to students—yet, this issue was not categorized as especially important by students. Conversely, despite its absence from the literature we reviewed, teacher interviewees described student responsiveness (or the lack thereof) as an important signal of respect and openness and grouped it among the three most important indicators of TSR. Because of its import in the literature and to teachers, we retained this indicator to potentially develop into an item. Thus, this step provides an informative source of complementary (or counterbalancing) data to the academic scholarship gleaned from Step 1.

Step 3: Synthesizing the Literature Review With Interview–Focus Group Data

The third step of this scale development process represents an initial attempt to reconcile differences that emerge between academic and lay conceptualizations of the construct in question. Specifically, the goal of this step is to provide a full conception of the construct with which both parties are likely to agree. From the literature review and the interview–focus group data, survey designers can develop a comprehensive list of indicators for their construct (from which they will develop initial items in Step 4). The merging of these sources of data is straightforward when prior literature and respondents agree on particular indicators. When they agree conceptually but describe the indicators in different ways, survey designers can use the vocabulary of their respondents. At this stage, when an indicator is mentioned from one source but not the other, most researchers will want to retain the indicator for the time being—later steps in the process will provide checks to see whether items that reflect that indicator seem appropriate.

We integrated the indicators mentioned by our participants and the literature in two ways. In some instances, academics and potential respondents agreed on the importance of certain aspects of TSR but described them differently. For example, "teacher support" is a prominent indicator of TSR in the literature (Goodenow, 1993; Murdock, Anderman, & Hodge, 2000). However, students actually described qualities like teachers "who took an interest in their personal lives." We noted this terminology in preparation for wording our items in Step 4. In other instances, we had to refine our conception of what the indicator represented. For example, the literature and the students noted the importance of teachers that hold high expectations that are individualized for students (Wentzel, 2002). However, teachers also reported students' expectations of them as important, but the key issue was teachers wanting students to hold realistic expectations of their role in the classroom—a crucial distinction from high expectations.

Step 4: Developing Items

After synthesizing their lists, survey designers can write preliminary items. The goal of this step is to develop items that adequately represent the indicators from Step 3 while using terminology that is meaningful to potential respondents (from Step 2). Two challenges predominate during this phase of scale development. The first challenge lies in determining the number of items to generate. For large-grain constructs, having an item that corresponds to each indicator may be impossible. (Failing to represent each indicator may chafe against researchers' instincts. However, even 100 items on a narrowly defined construct only asks a *sample* of all the relevant information that might be asked about that construct.) Deciding on the number of items ultimately rests with the professional judgment of researchers and (as we illustrate below) their facility in developing items that represent larger portions of the whole construct. However, the conservative path is to develop more items than are needed for the final scale (e.g., perhaps developing 15 potential items in the hopes of ultimately developing an eight-item scale). It may also help to check these items against the Venn diagrams developed in Step 1.

The second challenge of this step lies in actually wording each item. Much of the guidance on wording items in the typical survey construction approaches describes how designers should use clear, unambiguous language; guard against bias; ensure that the item stems cohere with the response anchors; be wary of offending respondents when asking for sensitive information; avoid double-barreled items; and so on. Although these are important reminders, they are probably intuitive for many survey designers. Conversely, a wide array of equally important issues appears to be less well-known among survey designers. Table 1 illustrates five such issues as well as ideas for addressing these issues. The goal of the table is to illustrate (rather than exhaustively document) some of the many issues that arise in survey design for which survey designers cannot rely on instinct. These issues have been studied empirically, and several best practices have emerged. We are agnostic as to whether survey designers familiarize themselves with these practices directly or familiarize themselves with a knowledgeable colleague instead. However, as the art of crafting items is increasingly becomes a science, new scales need to reflect the field's collective knowledge of this empirical work. In those instances where the evidence does not point to a clear right answer, familiarity with this literature will augment survey designers' awareness the trade-offs involved in different decisions.

In creating items for our TSR measure, we were concerned not only about the large-grain nature of our construct but also about the valence of different items. Our master list of indicators produced 14 topics—a sizable number, particularly for seventh grade attention spans. Thus, we clustered certain indicators (essentially creating subconstructs) and developed items to address those clusters. For example, rather than developing three separate items to address teachers' enthusiasm, warmth, and communication, we developed one item to address encouragement. To address the

Table 1
Lesser Known Best Practices in Item Development

Best practice	Explanation	References
Avoid using reverse-scored items	In theory reverse-scored items are a clever idea to "keep respondents honest" and prevent them from responding to all questions in the same way. In practice, theoretically ostensible opposites are frequently not arrayed on a continuum (i.e., they are multidimensional) and reverse-scored items diminish scale reliability.	(Benson & Hocevar, 1985; Cacioppo & Berntson, 1994; Swain et al., 2008)
Use at least 5–7 response anchors	For most respondent populations, 5-point response anchors for unipolar items (that range from a conceptual 0 point to infinity) and 7-point response anchors for bipolar items (that conceptually range from negative infinity to infinity) will work well.	(Krosnick, 1999; Weng, 2004)
Avoid agree–disagree response anchors	Asking respondents to rate their level of agreement to different statements is a cognitively demanding task that increases respondent error and reduces respondent effort in many cases.	(Fowler, 2009; Krosnick, 1999)
Label each response anchor with a construct-specific verbal label; avoid numeric labels	Labeling each response anchor enhances reliability. Using construct-specific anchors (e.g., *not at all/slightly/somewhat/quite/extremely interested* if "interest" is the construct in question) should help reinforce respondents focus on the core construct under investigation. Because numbers have implicit meaning for many participants—which may conflict with the verbal response anchors—they should be avoided.	(Tourangeau, Rips, & Rasinski, 2000)
Strive to ensure that every part of every question applies to every respondent	If parts of individual items make false assumptions about respondents, you will introduce error or missing responses into your data. For example, "How frequently do you see your family doctor?" assumes that respondents (a) have a family doctor, (b) see the doctor from time to time, and (c) it implies but does not clarify that "seeing" the doctor occurs for medical reasons (as opposed to for a weekly tennis match). Instead, survey designers can use branching items to filter respondents toward applicable items (e.g., if undergraduate seniors need to take an extra section of the survey, ask respondents for their class year and direct all nonseniors toward the next appropriate part of the survey).	(Bradburn, Sudman, & Wansink, 2004; Dillman et al., 2009)

issue of the valence of our items, we knew to avoid reverse scored items—especially for younger respondents (see Benson & Hocevar, 1985; Swain, Weathers, & Niedrich, 2008). Yet, we also knew from the TSR and attitude literatures that positive and negative behaviors and attitudes can impact relationships in different but important ways (Birch & Ladd, 1997). Thus, we drafted one set of items that served as indicators of positive TSR (e.g., "How friendly is Teacher X toward you?") and a separate set to serve as negative TSR indicators (e.g., "How often does Teacher X make you feel upset?").

Step 5: Expert Validation

With a list of potential items in hand, survey designers can return their focus to their academic audience. The expert validation step allows survey designers to collect data that establishes the construct relevance of individual items and double checks for key omitted indicators. This process can also provide information on item clarity, language complexity, and other item-level concerns researchers may have. Toward this end, survey designers can identify experts in the field (e.g., authors of relevant publications from the literature review) and invite them to judge how well a set of items represents a particular construct. Those judges then complete a survey in which they read the survey designers' definition of the construct; rate each potential item on construct relevance, clarity, or other characteristics of concern to the survey designer; write-in additional comments about individual items; and identify any important indicators that they perceive to be underrepresented or absent. (See the Appendix for a template of an expert review.) This process also offers designers the chance to quantify the content validity of their scale—see McKenzie, Wood, Kotecki, Clark, and Brey (1999) and Rubio, Berg-Weger, Tebb, Lee, and Rauch (2003) for examples of validation processes with corresponding content validity statistics. Another useful technique for designers who are creating multiple related scales is to ask experts to match items to the construct they belong to (or to an "other" category)—this process provides an early indication of how well the items manifest discriminant validity (Hinkin, 1995). See Veneziano and Hooper (1997) for guidance on the number of experts participants needed for this technique.

Because we needed our TSR items to capture a large-grain construct from the perspective of teachers and students, we focused our expert validation process on assessing construct relevance and ensuring that we had not omitted key indicators. Because we were developing these scales for use with teachers and students, we developed separate expert lists of scholars who were familiar with the literature on teachers and middle/high school students, respectively. Each expert received an invitation to participate and a survey packet. The resultant data illustrated several issues that helped us evaluate the clarity and content of our items. For instance, experts noted that some items, which were focused on in-class time (i.e., "How encouraging is Teacher X of your efforts in the classroom?"), were important to TSR regardless of whether they occurred in class or not. For these items, experts suggested that focusing exclusively on classroom time might misrepresent the TSR for some teachers and students. Our judges also encouraged us to better balance the number of social versus instructional indicators of TSR.

Step 6: Cognitive Pretesting

The penultimate litmus test, before conducting a larger scale pilot study, is to learn how potential respondents understand and respond to each item. The practice of "cognitive pretesting" or "cognitive interviewing" provides a structured approach for learning how respondents interpret items (Presser et al., 2004; Willis, 2005). Though specific approaches differ, the core of this technique usually entails the survey designer to interview potential respondents and ask them (a) to repeat the question in their own words—sometimes without repeating any words from the question itself and (b) to think out loud by reporting every thought they have as they answer the question. During or at the end of the interview the survey designer usually asks follow-up, probing questions to clarify how respondents understand each item. See Karabenick et al. (2007) for a detailed illustration of this process.

Two aspects of this technique are important to recognize. First, cognitive pretesting is strange and unnatural for most respondents—especially having to verbalize one's inner monologue to the outside world. Thus, many researchers begin their cognitive pretesting interviews with a practice item that allows respondents to practice and receive feedback on this unusual process. Second, this technique can also lead survey designers to overthink their items (Willis, 2005). In other words, how respondents interpret and answer questions during a survey is sometimes incongruent with their interpretations when they fixate on an item for a protracted period of time during a cognitive pretesting session. Thus, experts often advise identifying clear trends from multiple respondents about a potentially problematic item before making changes (Willis, 2005).

In our TSR survey, this cognitive pretesting technique illuminated three types of potential problems: ambiguity of meaning, overly challenging vocabulary, and ambiguity of situation. First, students noted a problematic ambiguity with "How often does Teacher X make you feel upset?" They observed that students may feel upset if they dislike their teacher or if they are invested in the class and their teacher gives them a disappointing grade. Thus, the item was unlikely to be the unequivocal negative indicator we had originally envisioned. Second, the item that asked participants, "How likely would Student X be to recommend you as a teacher to another student?/How likely would Teacher X be to recommend you as a student to another teacher?" used a word ("recommend") that some students did not fully understand. Third, for this same item, teachers and students felt that they needed more contextual knowledge to generate an answer (e.g., the personalities of the people involved and the subject matter of the class).

Pilot Testing

Despite the best efforts of survey designers and rigorous adherence to the previous six steps, some items may remain problematic. However, at this stage most problems are hard to detect without data from a larger sample. Thus, the goal of pilot testing is to administer the scale to a larger population of participants to test how items function within the scale and to determine how the scale functions relative to other measures. Fortunately, as described earlier, much has been written on the "pilot testing–analysis–item selection" procedures that comprise the final stages

of scale construction. Usually, researchers iterate through pilot testing until they have a core group of items that function well (i.e., a reliable, cohesive group of items that manifest construct validity in at least a couple different ways) on a sample that closely resembles their population of interest.

Our own pilot testing data identified some problematic items that we refined to make them more applicable to a broader array of teachers and students. For instance, student responses to, "How many times have you expressed anger toward Teacher X in the past week?" indicated that these students rarely expressed anger toward their teacher. To increase the variability, we reworded the item from expressing to feeling anger: "How angry does Teacher X make you feel during class?" In other instances, our pilot data illuminated items with substantial face validity that simply did not correlate highly with other items and had to be revised substantially or removed.

These six steps are clearly not a panacea for developing perfect items and, thus, cannot obviate the need for pilot testing. However, they did help us address several important problems early in the survey design process that multiple iterations of pilot testing may not have uncovered. For example, if we had finessed the review of the literature and simply relied on our existing knowledge, we would not have distinguished subtleties such as teacher caring as a personality trait falling outside of our construct of interest but the act of a teacher caring about a particular student as being germane to our construct of interest. Without Steps 2 and 3 we would have missed important indicators such as student responsiveness and students having reasonable expectations for their teachers as important facets of the relationship. Ignorance of the best practices in Step 4 would almost certainly have introduced more respondent error into specific items, for example, we might have been tempted to use items with agree–disagree response anchors that we found in some prior measures without adapting them. Step 5 is especially important in light of the typical approaches to survey construction. Through the typical approach, very large item pools are often constructed initially and then items are pruned through pilot tests (often when they prove problematic for the factor structure of the scale). However, this pruning can result in the omission of critical aspects of the construct. The expert testing that we conducted helped us to realize that we were in danger of underrepresenting the role of student learning in teacher–student relationships. Finally, Step 6 helped us realize that, despite our best efforts, we had still included some vocabulary that was too difficult for some of our respondents. Again, this important information could easily go undetected through pilot testing.

Caveats and Concluding Thoughts

Many topics pivotal to the successful administration of a survey fall outside our focus on scale development. Representative samples, maximizing return rates, use of incentives, layout–formatting of surveys, and modality (e.g., paper and pencil vs. web) are all critical topics reviewed elsewhere (e.g., Dillman, Smyth, & Christian, 2009). Moreover, we are not so naïve as to think that our six steps represent the perfect prepilot testing procedure; our process will likely be refined as researchers experiment with and adapt these steps. In the meantime, we are confident that psychologists who use these steps will identify and remediate more issues with their scales than those who simply make up items and pilot test them—an approach we suspect is too frequently used within psychological and social science research.

References

Ang, R. P. (2005). Development and validation of the Teacher-Student Relationship Inventory using exploratory and confirmatory factor analysis. *Journal of Experimental Education, 74,* 55–73. doi:10.3200/JEXE.74.1.55-74

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitudes scales for elementary school children. *Journal of Educational Measurement, 22,* 231–240. doi:10.1111/j.1745-3984.1985.tb01061.x

Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology, 35,* 61–79. doi:10.1016/S0022-4405(96)00029-5

Birch, S. H., & Ladd, G. W. (1998). Children's interpersonal behaviors and the teacher-child relationship. *Developmental Psychology, 34,* 934. doi:10.1037/0012-1649.34.5.934

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design – for market research, political polls, and social and health questionnaires* (Rev. 1st ed.). San Francisco, CA: Jossey-Bass.

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115,* 401–423. doi:10.1037/0033-2909.115.3.401

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319. doi:10.1037/1040-3590.7.3.309

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56,* 754–761. doi:10.1037/0022-006X.56.5.754

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.). Hoboken, NJ: Wiley.

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage.

Goodenow, C. (1993). Classroom belonging among early adolescent students: Relationships to motivation and achievement. *The Journal of Early Adolescence, 13,* 21–43. doi:10.1177/0272431693013001002

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21,* 967–988. doi:10.1177/014920639502100509

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., . . . Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42,* 139–151. doi:10.1080/00461520701416231

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537–567. doi:10.1146/annurev.psych.50.1.537

McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling.* Beverly Hills, CA: Sage.

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.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. doi:10.1037/0003-066X.50.9.741

Murdock, T. B., Anderman, L. H., & Hodge, S. A. (2000). Middle-grade predictors of students' motivation and behavior in high school. *Journal of Adolescent Research, 15,* 327–351. doi:10.1177/0743558400153002

Pianta, R. C. (1999). *Enhancing relationships between children and teachers*: American Psychological Association. doi:10.1037/10314-000

Pianta, R. C. (2001). *Student-Teacher Relationship Scale (STRS): Professional manual.* Lutz, FL: Psychological Assessment Resources.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). *Methods for testing and evaluating survey questionnaires.* Hoboken, NJ: Wiley. doi:10.1002/0471654728

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.

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass, 2,* 414–433. doi:10.1111/j.1751-9004.2007.00044.x

Simms, L. J., & Watson, D. (2007). The construct validation approach to personality scale construction. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 240–258). New York, NY: Guilford Press.

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45,* 116–131. doi:10.1509/jmkr.45.1.116

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

Veneziano, L., & Hooper, J. (1997). Research notes. A method for quantifying content validity of health-related questionnaires. *American Journal of Health Behavior, 21,* 67–70.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64,* 956–972. doi:10.1177/0013164404268674

Wentzel, K. R. (1997). Student motivation in middle school: The role of perceived pedagogical caring. *Journal of Educational Psychology, 89,* 411–419. doi:10.1037/0022-0663.89.3.411

Wentzel, K. R. (2002). Are effective teachers like good parents? Teaching styles and student adjustment in early adolescence. *Child Development, 73,* 287–301. doi:10.1111/1467-8624.00406

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

Appendix

Expert Review Template

Sample Instructions

Thanks again for agreeing to participate in our expert review of the items on the _____ *Scale* that we are developing. Below is a description of the larger research project, the construct definitions, and then a list of questions about each of the items on the survey. Please begin by familiarizing yourself with this background information and the construct definitions, and then review the specific instructions for completing the content validation.

I. Research project: [Describe relevant details of the research project possibly including: what your research questions are, who you are sampling, and how the survey fits into the research project.]
II. Construct definition: [Provide a clear definition of your construct.]
III. Clarity: In this section we would like to know how comprehensible each item is for our anticipated respondent population. Please rate how understandable each of the following items is by using the scales below. If you have ideas for how to clarify the meaning of an item please note your thoughts beneath each item.

1. Item 1 [including response anchors]

Not at all understandable	*Slightly understandable*	*Somewhat understandable*	*Quite understandable*	*Extremely understandable*

Suggestions: _____

2. Item 2 [including response anchors]

Not at all understandable	*Slightly understandable*	*Somewhat understandable*	*Quite understandable*	*Extremely understandable*

Suggestions: _____

. . . and so on.

IV. Item means: In this section we would like your help to anticipate which of our items will produce an adequate range of means. Please indicate what you think the average (mean) response for each item will be given our target respondents.
1. Item 1 [provide the actual verbal response anchors that you plan to use for each item]
2. Item 2 [provide the actual verbal response anchors that you plan to use for each item]

. . . and so on.

V. Relevance: In this section we would like to know how central each item is to our construct of interest. Please rate the relevance of each item to the construct of _____.
1. Item 1 [including response anchors]

Not at all relevant	Slightly relevant	Somewhat relevant	Quite relevant	Extremely relevant

2. Item 2 [including response anchors]

Not at all relevant	*Slightly relevant*	*Somewhat relevant*	*Quite relevant*	*Extremely relevant*

. . . and so on.

Next, please think about all the items as a whole for a moment. We hope this survey scale fairly represents the entire construct without ignoring important features of the construct. Please indicate any aspects or characteristics that you feel are important parts of this construct which are not represented or are inadequately represented by this survey scale.

1. _____
2. _____
3. _____

Received March 31, 2011
Revision received August 31, 2011
Accepted September 1, 2011 ■