

Establishing survey validity: A practical guide

William W. Cobern ^{1,*}, Betty AJ Adams ²

¹The George G. Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI, USA

ARTICLE HISTORY

Received: May 22, 2020

Accepted: Aug. 15, 2020

KEYWORDS

Pretesting,
Cognitive interviews,
Reliability,
Research Methods,
Survey Methods,
Validity

Abstract: What follows is a practical guide for establishing the validity of a survey for research purposes. The motivation for providing this guide is our observation that researchers, not necessarily being survey researchers per se, but wanting to use a survey method, lack a concise resource on validity. There is far more to know about surveys and survey construction than what this guide provides; and this guide should only be used as a starting point. However, for the needs of many researchers, this guide provides sufficient, basic information on survey validity. The guide, furthermore, includes references to important handbooks for researchers needing further information.

1. INTRODUCTION

We have written this practical guide because of a dispute that arose between two faculty members and a student. The faculty members criticized the student for having insufficiently established the validity of a survey she had created. As the student was working under our supervision, the criticism was surprising. On the other hand, we quickly realized that the situation constituted a proverbial “teachable moment.” Even though the student had taken a course on survey development and we had discussed the methodology, we realized that neither students nor faculty had a practical guide on how to establish survey validity, or what that even means. This document is an attempt to fill that need.^{†,‡,§} These are not survey researchers per se, but researchers who on occasion need to develop a survey for the purposes of their research interests.

CONTACT: William W. Cobern ✉ bill.cobern@wmich.edu 📍 The George G. Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI, USA

[†] This guide does not address the purposes for survey research. The assumption of this guide is that the researcher has already made the decision to use a survey. This guide is solely about the production of a valid survey for research purposes.

[‡] Boateng et al. (2018) offers a similar practical guide but from a different perspective with somewhat different coverage.

[§] Much of what is in this practical guide can also be applied to the development and validation of interview protocols.

At the start it is important to distinguish between surveys and tests, though in fact much of this practical guide is also relevant to test construction. Tests and surveys have much in common, indeed, sometimes it is difficult to tell the difference. For example, is the *Student Understanding of Science and Scientific Inquiry* (SUSSI) a test or a survey? Is the *Views of Nature of Science Questionnaire* (VNOS) a test or a survey? Is the *Measure of Acceptance of the Theory of Evolution* (MATE) a test or survey? Is the PEW instrument for assessing public knowledge of science a test of knowledge or a survey of knowledge? Could be either. For the purposes of this practical guide, we make the following distinction. Surveys (or questionnaires**) typically collect information about attitudes or opinions, can also be used to survey knowledge, but are typically not associated with instructional settings. On the other hand, tests are almost always about knowledge or skills and, unlike surveys, tests generally are associated with instruction. This is not a hard and fast distinction, however, so in this practical guide we will use examples that some people may think of as tests; it makes no difference to the procedures we present.

This practical guide is purposefully simple as the objective is to provide practical guidance on a few basic things that all researchers should observe for establishing survey validity. Furthermore, one can think of survey construction as serving one or two purposes. Researchers may construct survey instruments because they need an instrument to collect data with respect to their specific research interests. The survey is not the focus of the research but a tool, an artifact of conducting research. Other people may decide to use the researcher's instrument as they see fit, though it was not the researcher's intention to provide a new instrument for other researchers to use. For example, Barbara Greene studies cognitive engagement and for this purpose she and her colleagues have developed a number of survey-type instruments. She writes about getting regular requests from others wishing to use her cognitive engagement scales, which came as a surprise to her group as they developed the scales for their own research purposes (Greene, 2015). They were not in the business of developing instruments for general research use. On the other hand, some research is specifically about survey construction of which there are many examples including Lamb, Annetta, Meldrum, and Vallett (2012), Luo, Wang, Liu, and Zhou (2019), Staus, Lesseig, Lamb, Falk, and Dierking (2019).

Survey development can involve powerful statistical techniques such as Item Response Theory (Baker, 2001) or Rasch Modelling (Boone, 2016). One is more likely to see these techniques used when a survey is developed for broad use. These techniques are less common when a survey instrument is developed as an internal artifact for conducting specific research. Perhaps more often one will see researchers employ factor analyses as part of survey development. This practical guide does not address either Rasch Modelling or Item Response Theory, and only mentions factor analysis in passing. Our focus is on the development of narrowly focused surveys designed for the research a person wishes to pursue, and not on the development of a survey for others to use. Of course, for whatever reason someone produces a survey, as noted above, that survey is likely to get used by others regardless of the originator's intention for the survey.

Surveys serve a broad range of purposes. Some are simply seeking factual or demographic information. We may want to know the age range across a group of people. We may wish to ask students enrolled in a particular course what their majors are. We might be interested in how a group of people prioritizes a set of unambiguous entities. On the other hand, we might be interested in using surveys to gauge far more complex constructs such as attitudes, behaviors, or cognitive engagement. The latter are much more difficult to develop and validate than are the former.

** We do not think that there is anything in the literature that provides a strong rationale for distinguishing between surveys and questionnaires. For all practical purposes, there is no difference. The research literature, however, typically uses the word survey.

Whether using sophisticated methods such as Rasch Modelling, Item Response Theory, or factor analysis, or more basic methods, whether developing a simple survey or a rather complex one, every researcher begins with three questions that are not necessarily easy to answer:^{††}

- 1) What is it that I want to learn from the responses on my instrument?
- 2) What assurance can I have that my respondents understand what I am asking?
- 3) How can I be reasonably sure that the responses my respondents give to my items will be the same responses they give to the same items two weeks later?

The first and second questions are about instrument validity, and the third question is about instrument reliability.

2. EVIDENCE SUPPORTING VALIDITY

What is this idea of validity? Here is an example to help illustrate the general idea of validity. If you give students a set of questions having to do with their interest in science and they consistently respond about their interests in the arts, there is a problem. The questions prompted consistent^{‡‡} responses but the responses are not about the information you were seeking. Somehow, the questions give the respondents the wrong idea that you wanted to know about their interest in the arts when what you wanted was to know about their interest in the sciences. Your questions are not valid with respect to the information you are trying to get. A test item or survey item (and this applies to interview items as well) has validity if the reader of the item understands the item as intended by the item's creator. As stated in the 2018 *Palgrave Handbook of Survey Research* (Vannette & Krosnick, 2018):

An important aspect of validity is that the survey is designed in such a way as to minimize respondent error. Respondent error has to do with the respondent responding to an item in some way that is different from the researcher's intention. (Krosnick, 2018, p. 95)

Validity is an evidence-based argument. The researcher provides evidence that the instrument is valid with respect to its intended purpose and audience. According to the 2014 *Standards for Educational and Psychological Testing*,

Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use. (AERA, APA, NCME, 2014, p. 11)

At least since the 1999 *Standards* edition, measurement experts in education and psychology have ceased referring to distinct types of validity (e.g., content or construct validity)^{§§}, preferring to view validity as a unitary concept represented by the “degree to which all accumulated evidence supports the intended interpretation of test scores for the proposed use” (AERA, APA, NCME, 2014, p. 14). Moreover, as one might expect, there are various sources and types of evidence:

That might be used in evaluating the validity of a proposed interpretation of test scores for a particular use. These sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. (AERA, APA, NCME, 2014, p. 13-14)

^{††} Our epistemological perspective is that survey development and validation are processes that need to proceed hand-in-hand. We do not consider it wise for the researcher to separate these processes into a sequence of development first followed by validation

^{‡‡} Consistency has to do with reliability and is discussed later.

^{§§} See Ruel et al. (2016) for an example from sociology of researchers retaining the old system.

Furthermore, “the wide variety of tests and circumstances makes it natural that some types of evidence will be especially critical in a given case, whereas other types will be less useful” (AERA, APA, NCME, 2014, p. 12).

It is beyond the scope of this practical guide to present much detail on the various types of evidence that can be used in support of validity. For that purpose, readers should consult authoritative documents such as the 2014 *Standards for Educational and Psychological Testing* or the 2018 *Palgrave Handbook of Survey Research*. However, for practical purposes, there are two areas of importance for establishing evidence of validity: a validated model that provides the basis for an instrument, and the items composing an instrument.

2.1. Foundational Model

A valid survey requires a theoretical model of what it is the researcher wants to find out by having people respond to survey items. The foundational model answers the question: *What is it that I want to learn from the responses on my instrument?* Answering this question involves obtaining or building a validated, theoretical model for what the researcher wants to know. Beware of the temptation just to write items straightaway. This happens far too many times where the researcher completely skips the idea of theoretical model building and jumps directly into writing items (or questions).^{***} These are items simply coming to one’s mind but lacking theoretical foundation. Such items are *ad hoc*, and an instrument built on *ad hoc* items is not a research-worthy instrument. There is already a validity issue because there is no foundation for the survey. The first line of validation evidence for survey items is the foundational model.

While there probably are many ways to develop a foundational model, these ways certainly include theory-driven model development, statistically-derived model development, and grounded theory model development. Theory-driven model development is a top-down approach in contrast to the bottom-up approach of statistically-derived model development and grounded theory model development. Bottom-up model development is essential when the researcher has no *a priori* model or theory on which to build a survey. In that situation, the model has to be built inductively from data collected from the type of people who would ultimately become subjects of research where the survey is used, or possibly built inductively from expert opinion. Bottom-up model development oftentimes involves a combination of grounded theory and statistical analysis. For example, let’s say you are interested in the goals that college faculty have for chemistry lab instruction and you would like to survey a large number of college chemistry faculty to determine what goals are most frequent. Bruck and Towns (2013) developed such a survey that began with a grounded theory approach. Initially, the researchers collected qualitative data from interviews with college chemistry faculty on the goals they had for chemistry lab instruction (Bruck, Towns, & Bretz, 2010). Subsequently,

An initial pool of survey items was developed from findings of the qualitative study. Questions constructed from key interview themes asked respondents to identify the frequency of certain laboratory practices, such as conducting error analyses or writing formal laboratory reports. (Bruck & Towns, 2013, p. 686)

When these researchers say that they developed an initial pool of items drawing from the findings of their qualitative study, they are essentially describing a grounded theory approach. They are “on the ground” with college chemistry faculty finding out directly from them what their goals are. However, this data has no structure; it represents no model. To create a foundational model that provides structure for a survey based on the ideas coming directly from the faculty, the researchers turned to statistical methods. The researchers drafted a survey using

^{***} People use the terms ‘item’ and ‘question’ interchangeably with regard to surveys. ‘Item’ is the more general term but items on a survey are all questions in that each item represents a request for information whether it is, for example, one’s birthday or one’s opinion registered on a Likert scale.

these items that they then distributed to a large number of chemistry faculty. They subjected the resulting data to statistical procedures (correlation tables, Cronbach's α , Kaiser-Meyer-Olkin tests, and factor analysis) resulting in a seven-factor model:

| | |
|----------------------------------------------|----------------------------------------|
| Research Experience | Transferable Skills (Lab-Specific) |
| Group Work and Broader Communication Skills | Transferable Skills (Not Lab-Specific) |
| Error Analysis, Data Collection and Analysis | Laboratory Writing |
| Connection between Lab and Lecture | |

In the process, the researchers dropped items not fitting this model (i.e., those having low statistical value) resulting in the 29-item *Faculty Goals for Undergraduate Chemistry Laboratory Survey*, a survey for which the foundational model was derived bottom-up using a combination of grounded theory and statistical methods. The validity lines of evidence include the initial qualitative data gathered from interviews and the subsequent statistical analyses of data. For an instrument derived from a combination of grounded theory and statistical methodology, the building and validation of the model and the instrument are intertwined. They go hand-in-hand.

The development of a theoretically derived foundational model is much different, though the question remains the same: *What is it that I want to learn from the responses on my instrument?* The difference is that the researcher already has a model or theory on which to base the instrument; hence, the development approach is top-down. The survey is derived deductively from the model. Such models can come from the literature (which is often the case) or researchers construct the model by drawing from the literature. In either case, the connection to the literature validates the model. Moreover, it is possible for researchers to invent a model to suit their philosophical positions and research interests. Our first example comes from research conducted by the first author and is an example of a model drawn from the literature (Cobern, 2000).

Cobern, Gibson, and Underwood (1999) and Cobern (2000) reported investigations of how students conceptualize nature, that is, the natural world. The studies had to do with the extent to which students voluntarily introduce science into their explanations about nature. These were interview studies rather than survey studies; but the theoretical modeling would have been the same had Cobern decided to collect data using a survey. A wide-ranging review of the literature led to a model involving four categories of description along with a set of disparate adjectives that could be used to represent each category description (see [Table 1](#)).

This model represents what Cobern wanted to learn from the study. He wanted to learn the various ways in which students might describe nature, and for reasons described in the published papers, he based the interview protocol on this *a priori*, theoretical model. Basing the interview protocol on the theoretical model provides the first line of validity evidence. The same would be true if he had decided to use a survey method. Deriving the survey from a literature-validated model^{†††} provides the first line of validity evidence for the survey.

^{†††} The literature-based validation of a model does not mean that one particular model is the only one a researcher could validate from literature. Undoubtedly, in most situations, literature can validate a number of different models. Therefore, the onus is on researchers to explain why they built a particular model and on readers to judge that explanation.

Table 1. Modeling: what is nature? (Cobern, 2000, p. 22)

| | | | |
|-----------------------------------------------------------------------------------------------|--------------------------------------------------------|--------------------------------------------------------|-----------------------------------------------|
| <u>Epistemological Description:</u> (Reference to knowing about the natural world.) | confusing mysterious | unexplainable unpredictable | understandable predictable knowable |
| <u>Ontological Description:</u> (Reference to what the natural world is like.) | material matter living complex orderly beautiful | dangerous chaotic diverse powerful changeable | holy sacred spiritual unchangeable pure |
| <u>Emotional Description:</u> (Reference to how one feels about the natural world.) | peaceful | frightening | "just there" |
| <u>Status Description:</u> (Reference to what the natural world is like now.) | "full of resources" endangered | exploited polluted | doomed restorable |

It is important to understand that the above examples involve categories that subsume items or interview questions. Respondents address the items, not the categories. For example, the Bruck and Towns (2013) survey does not explicitly ask respondents about “research experience,” which is one of their categories. “Research experience” is too ambiguous a term (see section on item clarity below) to ask about it explicitly. Rather, respondents see a set of clearly stated items that according to the researchers’ model represents “Research Experience.” Thus, respondents do not need to understand the construct; they only need to understand the language of the items in which the construct is expressed. A consequence of such modeling is that the internal consistency of categories needs to be checked every time the instrument is used. Researchers should not assume “once validated, always validated.”

The Cobern (2000) model was constructed *from* the literature; however, in other cases, a top-down model may be found *directly* in the literature. In other words, the model is not derived from the literature but is literally borrowed from the literature. For example, Haryani, Cobern, and Pleasants (2019) investigated Indonesian teachers prioritizing of selected curriculum objectives. Their national Ministry of Education establishes the Indonesian curriculum and it is incumbent upon all Indonesian teachers to know and follow this official curriculum. Haryani et al. (2019) was specifically interested in the new addition of 21st Century Learning Skill objectives to the curriculum (creativity and innovation, critical thinking, problem-solving, collaboration, and communication skills), and how teachers prioritized these new objectives. The model for the research survey (Table 2 below) came directly from the official curriculum. Basing the survey items on this theoretical model read from the literature (i.e., the official curriculum) provided the first line of validity evidence.

Summarizing this section, establishing the validity of an instrument begins with clearly answering this question: what is it that I want to learn from the responses on my instrument? Answering this question begins with having a validated, theoretical model (a foundational model) for what the researcher wants to know. The next section is about constructing a survey based on a model: item fit, instrument length, item format, item discrimination, item clarity, order of items, and item effectiveness.

Table 2. *Modeling: teacher C13-curriculum priorities*

| The C13 Curriculum Content | Outcomes |
|-----------------------------------------|---------------------------|
| Traditional C13 content | Science Content |
| Recent C13 additions | Science Processes |
| 21 st Century Learning Skill | Creativity and Innovation |
| 21 st Century Learning Skill | Critical Thinking |
| 21 st Century Learning Skill | Problem Solving |
| 21 st Century Learning Skill | Collaboration |
| 21 st Century Learning Skill | Communication Skills |
| C13 Irrelevant content | History of Science |
| C13 Irrelevant content | Writing Skills |
| Participant demographics ^{###} | Gender, school type |

2.2. Fitting Items to The Model

As noted earlier, sometimes the researcher is tempted to start instrument development by simply writing items as they come to mind. That temptation needs to be avoided by giving due attention to first building a model or acquiring one. With a model in hand to inform the development of the instrument, the researcher can either write original items or find useful items in the literature to use as-is or revised, or build an instrument from a combination of both. As items are gathered, they need to be fitted to the model. The model serves as a device for disciplining the selection of items. Furthermore, the fit should be validated by persons external to the instrument development process. In other words, the researcher should have a few, knowledgeable people check items for fit with the model.

Instrument length: Selecting items (or writing items) raises questions about the number of items, the wording of items, and item type. Regarding the number of items and thus the length of a survey, the rule of thumb is that shorter is better than longer. As noted by Krosnick (2018, p. 95), “the literature suggests that what goes quickly and easily for respondents also produces the most accurate data.” In other words, the threat to validity increases with instrument length.

Researchers need to minimize the length of a survey; but if a survey has to be long then precautions are needed because excessive length will very likely introduce response errors.^{§§§} For example, Nyutu, Cobern, and Pleasants (2020) needed student responses to 50 items in order to build a model (using a bottom-up approach) for their work on faculty goals for laboratory instruction. The researchers were concerned that students would not take the last items seriously given the length of the survey. To mitigate the potential problem, the researchers used five different forms of the survey where the item order was different on each form. By doing this, response errors in the last items would not be concentrated in the same items. This approach does not eliminate the problem but it at least eliminates the impact on specific items. Another approach would have been to use filter items toward the end of the survey. The researchers could have added one or two items toward the end that requested a specific response. For example, an item could have simply read, “For this item, select 3.” Thus, any survey that did not have a “3” for this item would have to be considered suspect. There are no perfect solutions when working long surveys but there are strategies, each with its advantages and disadvantages.

^{###} The inclusion of last three elements in this model, which are not 21st Century Learning Skills, is explained later.

^{§§§} For example, if a survey is very long then respondents may not pay attention to the last items because they have become tired of responding to so many items.

Of course, the best thing to do is to keep a survey short, and the model will help limit the number of items selected. However, researchers oftentimes want demographic data and this is where survey length can get out of hand. Note that the last entry in [Table 2](#) is participant demographics. The researchers specifically placed demographics in the model as a reminder to only ask for demographics that were important with respect to the rest of the model. For example, if the researcher does not have a good reason (that is, reasons relevant to teacher prioritizing of curriculum objectives) for asking teachers about their age, then the researcher should not ask for age. The researcher should only ask for demographics that are important to the study or for which the researcher has good reason to think could be important. Researcher discipline about demographic information helps keep survey length reasonable, bearing in mind that excessive survey length poses a threat to validity.

2.3. Item Format

The type of items to be used is another important question specific to survey development. Survey items frequently use Likert scales, which raises the question of how many points should be on a scale. Conventional wisdom is to use an odd number such as five or seven (Krosnick, 2018, p. 99). However, sometimes a researcher wants to avoid having respondents select a middle or “neutral” position, in which case the scale has to be an even number. Too few points or too many points threaten validity, and could either blur or exaggerate variation.

Survey items are oftentimes about information where the Likert format is not useful. Writing such items is fairly straightforward when the information is simple such as age. Asking how often somebody engages in activity can be trickier. For example, asking how often students watch YouTube videos has to begin with the assumption that students are unlikely to have a good idea of exactly how much time they spend per week watching YouTube videos. Hence, asking how many hours some spend watching YouTube each week is likely to return unreliable responses, mere guesses. Students will be more reliable approximating their viewing time given a choice of time intervals such as a) 0 to 5 hours per week, 6 to 10 hours per week etc. The challenge for the researcher is to create reasonable time intervals. While there are no guidelines or rules to help the researcher, the researcher can check the literature to see the kind of time intervals that have been used by other researchers and use that as a guide; or the researcher can create the intervals with respect to the needs of the research. By the latter we mean that the researcher decides reasonable magnitudes for the poles based on the nature of the research questions. Again, using YouTube viewing as an example, the researcher may decide that watching YouTube 10 hours a week would be a lot and that few students are likely to do that. On the other hand, the researcher might reason that most students would watch for at least an hour. Following this line of reasoning, the lower time interval might be 0 to 1 hour with the upper interval being 10 hours or more: a) 0-1 hrs, b) 2 to 5 hrs c) 6 to 10 hrs d) 10+ hrs. And as should be common practice, it is a good idea to have somebody outside of the research check the researcher's decision. For example, if the item is intended for students then the researcher should ask a few students about the item. For example, the researcher might ask the students if these are the time intervals they would use or if they would use different categories.

2.4. Item Discrimination

A common threat to validity comes from lack of discrimination. For example, if items, written to represent the model in [Table 2](#), simply ask what priority a teacher gives for each objective, the researcher could easily find that teachers give a high priority to all objectives, given that the official curriculum mandates all objectives. However, it is unreasonable to think that, even with a mandated curriculum, teachers would give every objective the same priority; thus, such a survey would fail to provide discrimination and the argument for validity weakened. Haryani et al. (2019) attempted to avoid this problem by using bipolar items that required the respondent

to compare objectives. For example, an item asked for “Critical Thinking” to be ranked with respect to “Problem Solving.” By this method, it was not possible for a respondent to give every objective the same priority. Discrimination was improved and thus validity was improved.

Another strategy for improving validity is to use distractor items. Distractor items represent elements that do not fit with the foundational model. If the survey is valid, respondents will reject the distractor items. Once again consulting Haryani et al. (2019), their model (Table 2) has two entries labeled irrelevant content. The survey built on this model contained distractor items that asked respondents to compare legitimate objectives with irrelevant content. The researchers obtain a further line of validation evidence if the respondents reject the distractor items as per the model.

2.5. Item Clarity

The lack of item clarity can potentially harm validity, and thus another line of validation evidence is that items are clearly written. There are resources that provide conventional wisdom on the wording of items. For example, Krosnick (2018, p. 100-101) suggests that items be simple and direct, containing no jargon or ambiguous words, or emotionally charged words. Items should not contain double clauses.^{****} He argues that it is better to avoid negations and important to avoid writing questions that lead the respondent in a particular direction. He suggests that it is a good idea for the researcher or researchers to read their items aloud before finalizing them because hearing an item can help one detect a lack of clarity.

Clarity of expression includes clarity of terms and concepts. The terms and concepts used in an item need to be ones with which respondents are reasonably conversant. Such clarity is rarely a problem for simple terms such as age. A response on age is never going to be exact but it is highly probable that what a respondent records for age will be within an error range of +/- 6 months. That error range is not going to be a problem for most education research. Haryani et al. (2019) used terms that came from the Indonesia national curriculum. These terms are more complicated than “age” and a person unfamiliar with the Indonesia national curriculum could easily misinterpret the terms. However, in a centralized education system where objectives are mandated, Haryani et al. (2019) reasonably assumed that Indonesian teachers in that system are conversant with terms found in that curriculum. On the other hand, researchers can quickly run into trouble if they use terms open to interpretation amongst potential respondents (Smyth, 2016). For example, a Likert item asking how often a teacher uses an inquiry approach to instruction will be subject to a wide range of teacher interpretations. A better approach would be to describe the teaching approach and then ask a how often the respondent might use this approach or one similar to it (see for example Cobern et al., 2014).^{†††}

Moreover, even apparently simple words can be potentially troublesome. Redline (2013), for example, found that a survey asking about “shoes” was open to various interpretations. Does the word shoes include boots? Or sandals? In a test of wording, Redline found that an item specifying the meaning of “shoes” returned different responses from an item that didn’t. The better approach is to break the question down into a set of specific questions, such as how many shoes do you have, how many boots do you have, have any sandals do you have, etc. The point is that when writing items, the researcher needs to explore ways of making sure that critical terms in an item will be understood as intended. Even small wording changes can change how respondents interpret and respond to an item. Cobern, Adams, Pleasants, Bentley and Kagumba (2019), for example, got substantially different survey results in a nature of science study when

^{****} Often referred to as ‘double-barreled’ items.

^{†††} If the researcher decides to use specific examples, such as specific examples of teaching approaches, then those examples need to be based on the theoretical model for the study. For example, see Haryani et al. (2019).

the wording of one item was changed. They also found that a change of item wording can have effects on written responses.

If researchers decide to create their own research instrument it's typically because nothing published meets their particular needs. Nevertheless, researchers are likely to find published surveys that are similar to what they need and it is wise to learn from such published efforts. For example, there are many science attitude surveys. Although none of these may meet a researcher's need, the researcher can still learn from published science attitude items. Moreover, while a perfectly applicable instrument may not be available, it is likely that there are existing questions specific to the interests of the researcher. This is particularly true for questions about demographic information. It is common practice for surveys to include questions about demographics, and example questions are easily available online (e.g., Bhat, 2019; Fryrear, 2016; Rosenberg, 2017). Because the effective wording of survey items is so critical to validity it only makes sense for researchers to learn from published research when writing new items, and to use existing items of known validity when possible.

2.6. Ordering of Items

Once the researcher has finalized a set of items, these items have to be ordered for effective presentation (Smyth, 2016). A researcher may be tempted to give little thought to the order in which items are presented in a survey but that would be a mistake. For example, unless there is a specific reason to group similar items together, grouping similar items runs the risk that the first items in the group will influence the responses to the later items in the group. Unless that is what the researcher wants, similar items need to be dispersed throughout a survey typically by randomly assigning position. The rule of thumb on survey length is that respondent attention wanes toward the end of a long survey. Therefore, any items considered critical are best placed towards the start of a long survey. Demographic questions are often listed at the end of a survey because in academia these items are typically less important than the content items, or at least require less deliberation by the respondent. Unless the research question is *how* demographics relate to the content answers (in which case you need both), better to lose demographic data than to lose data having to do with the main focus of the survey. Another possible criterion for ordering items is the amount of reflection a content item requires. Researchers may wish to place items requiring less reflection earlier in the survey so as to help ease the respondent into the survey. The point is that the importance of ordering items should not be overlooked; it is something the researcher should attend to before finalizing an instrument.

3. PRETESTING FOR ITEM EFFECTIVENESS

When potential respondents read an item, they need to understand the item as per the intention of the researcher. Item effectiveness is a matter of item validity. If a potential respondent does not understand the item as intended by the researcher then the respondent won't actually be responding to what the researcher intended to ask. The lines of evidence for item validity include what has been discussed above: model-based items, appropriate item format, item discrimination, item clarity, and item order. Nevertheless, items should always be pretested (Willis, 2016). Once researchers have finished the ordering of items, the items can be pretested as a whole instrument.

Pretesting begins with an external review of the items. The researcher should always have items read by persons who are similar to those for whom the survey is being constructed (the target population) or who are familiar with the target population. In addition, the researcher needs to have expert readers who are knowledgeable about the subject matter and can read items with respect to content (Dillman, Smyth, and Christian, 2014, p 249-250). The researcher needs to have a set of questions for the external reviewers to think about. For example, external reviewers might be asked:

- Having read the items, what do you think this survey is about?
- Do you think that subjects in our target population [stipulate that population] will have difficulties understanding any of these questions?
- Are there items that you suspect most respondents will answer the same way? In other words, are there items that you suspect will not return a range of responses? That is, most everyone will respond similarly.
- Are there any changes to items that you would recommend making? Changes that would make items more easily understood.
- Is any of the content wrong in your opinion?
- How much time do you think it will take a person in our target population [stipulate that population] to thoughtfully complete the survey?

A researcher might give external reviewers a copy of a survey along with these questions asking the reviewers to respond to the survey items and then to these questions. The researcher uses subsequent feedback for making adjustments to individual items and perhaps the survey as a whole. Or, this feedback could be obtained through interviews (see next section on cognitive interviewing) or focus groups. As noted earlier, the researcher may need to have two types of external reviewers: external reviewers who represent the target population and content expert external reviewers.

Pilot studies can also be useful for evaluating item effectiveness, though typically you would not conduct a pilot study prior to having an instrument externally reviewed. A pilot study involves having a sample from the target population take the survey. The researcher can check the pilot study data for the presence of seriously skewed item responses. Such items fail the objective of having items that discriminate amongst the respondents. Lack of correlation between items represents one kind of problem – weakening the targeted construct; strongly correlated items can mean another kind of problem, indicating too little difference between the items – again, little or no discrimination. If a survey contains filter or distractor items, these can be checked through a pilot study. If these items function as expected, the argument for validity is strengthened.

3.1. Pretesting via Cognitive Interviewing

As noted above, pretesting can also include “cognitive interviews,”

...an applied approach to identifying problems in survey questionnaires and related materials, with the goal of reducing the associated response errors. ... The cognitive interview is conducted using verbal probing techniques, as well as “think-aloud,” to elicit thinking about each question (Willis, 2018, p. 103).

Cognitive interviews are critical for surveys that are to include theoretical constructs because item validity rests on respondents understanding the construct intended by the researcher. The goal of interviewing is to determine the likely ways in which respondents from a target population will interpret constructs important to the research. Earlier we gave the example of how respondents can misunderstand the word “shoes.” Here is another example. If you ask a person if they own a car, how will they interpret “car”? Does car include small trucks and SUVs? Interviews with persons from the target population would give the researcher at least some insight into how broadly or how narrowly the concept of “car” is likely to be interpreted (Blair & Conrad, 2011).

If concepts such as “cars” and “shoes” are open to various interpretations, just think about the many ways that students or teachers might define the concepts of “teacher centeredness” and “student centeredness,” or in science specifically, the concept of “inquiry instruction.” A survey could ask science teachers, using a Likert scale, to what extent they think inquiry instruction is

effective; but the problem is that you couldn't be sure exactly what the teachers meant by inquiry.

Any time a researcher is considering the use of survey items that include potentially ambiguous concepts, cognitive interviews are critical. It is during an interview that the researcher can learn to what extent a “potentially” ambiguous concept is actually ambiguous. If it turns out that the concept is not ambiguous during an interview, then the researcher can go ahead. However, if the concepts are found to be ambiguous then different strategies are needed for item structure. An interview might yield a narrow range of meanings and in that case an item might indicate this range by, for example, putting a few clarifying terms in parentheses after the concept. Or, instead of writing a single item, the researcher could consider writing a set of items using the various terms uncovered during the interviews. However, the researcher could find it difficult to interpret the results from a set of items ranging around the researcher's intended concept.

Another possibility for dealing with potentially ambiguous concepts is to write scenarios or vignettes that (for the researcher's purposes) represent an intended concept. The idea is that a description of an event communicates more clearly than does a label for an event. For example, questions about a short description (or vignette) of an inquiry lesson (as per the researcher's definition of inquiry) should return more valid responses than merely asking a respondent about inquiry lessons where the definition of “inquiry” is left up to the respondent. Bear in mind that if vignettes or examples are to be used, these also need to be based on the foundational model for the research, otherwise validity is threatened.

Cognitive interviews are not without problems. Government survey labs in the USA make widespread use of cognitive interviews for evaluating public opinion surveys; however, Willis (2018, p. 104) notes that “it is unclear whether or not independent researchers testing the same questionnaire would reach the same conclusions.” Willis (2018, p. 104) further notes that little is known about “under what conditions are cognitive interviewing results stable and reliable, and what can researchers do to enhance those conditions.” Furthermore, all pretesting is influenced by sample size.

From the perspective of sample size, a problem's prevalence affects the number of pretest interviews needed to identify it. For example, if we conduct a specified number of cognitive interviews (n) and a particular problem (f) occurs with prevalence (π), what is the probability (P_f) that it will be observed at least once by the n^{th} pretest interview, i.e., at some point in a sample of size n ? The probability of observing a problem in the pretest sample depends on two factors: how often the problem occurs (π) and how likely it is to be detected when it does occur (d) (Blair & Conrad, 2011, p 640-641).

Blair and Conrad (2011, p 636) found that,

Multiple outcome measures showed a strong positive relationship between sample size and problem detection; serious problems that were not detected in small samples were consistently observed in larger samples.

Hence, the difficult question is how many interviews to conduct, because the more interviews one conducts, the less likely it is that the researcher will miss problems. Fortunately for most education researchers, the saturation rule can be used as a guide (Cobern & Adams, 2020; Seidman, 2006). This rule advises interviewing people until the researcher ceases to hear anything new. Admittedly, this rule does not guarantee that the researcher won't miss rare opinions but the researcher accepts this risk on the basis that rare misunderstandings of an item will not have a significant impact on the research. Moreover, finding a rare misunderstanding of an item does not necessarily suggest a corrective action. Consider the possibility that a researcher interviews as few as 10 people finding that nine of the 10 understand the item as written. Does the researcher change the item because of the one person out of 10 who

misunderstood the question? Probably not. Chances are if the researcher changes the item in light of that one person the other nine might then have difficulties (see Dillman et al., 2014, p. 248). The point is the researcher only needs to interview enough people to be assured that the item is generally understood. Changing an item requires that there be a general misunderstanding or perhaps a significant misunderstanding of an item on the part of several people – not just one.

3.2. In Summary

Surveys should always be pretested and if survey items include potentially ambiguous concepts then the researcher should use cognitive interviews to evaluate such concepts for ambiguity. Whether or not to rewrite an item based on the findings of cognitive interviews is a matter of researcher judgment. Typically, the researcher would not rewrite an item unless the interview findings indicated substantial potential for misinterpretation. If an item is to be rewritten, there are various approaches other than simply changing the words of the item. The researcher can consider using, for example, a set of questions rather than one, adding descriptions to clarify the potentially ambiguous concept, or employing illustrative vignettes in the place of terms.

4. PILOT TESTING FOR RELIABILITY

Having done all of the above in order to have a strong argument for survey validity, there remains the question of how reasonable it is that respondents' responses are stable. Put another way, if you ask respondents the same set of questions two weeks later, will they respond the same way? This stability is what reliability is about (AERA, APA, NCME, 2014, p. 33-47). Survey items are reliable to the extent that responses are stable. The responses don't change over short periods of time during which it is reasonable to assume that nothing has occurred to change respondent views.

Many researchers report Cronbach's alpha as a measure of instrument reliability. Following Taber (2018), we believe this to be a mistake. Cronbach's alpha indicates the internal consistency amongst the group of items. If you have a category, such as "Research Experience" referred to earlier, represented by a set of items, those items need to be highly correlated if they are to validly represent this category. The correlational strength can be gauged using Cronbach's alpha. However, internal consistency is not the same thing as stability over time, which is what reliability is. Hence, a better way to gauge reliability is to give same group of people the instrument twice and then calculate the correlation between two sets of responses (Multon, 2010). The standard benchmark for reliability is that the two episodes of taking the instrument correlate at 0.70 or better. The researcher must bear in mind that testing for reliability is sensitive to the size of the sample. The reliability test-retest will not be effective if the sample is too small. There is no hard and fast rule about how much time should separate the test and retest but conventional wisdom suggests a separation of 10 days to two weeks. There needs to be enough separation so that the first test has faded in the respondent's mind; but the separation cannot be too long because of the risk of intervening factors that would change respondent opinions recorded by the retest.

Finally, the data from a reliability test-retest should also be examined for validity. For example, factor internal consistency should be rechecked, response distributions for items should be rechecked, and the effectiveness of filter or distractor items should be checked.

5. CONCLUSION

As noted at the beginning, this document is a practical guide. There is far more to know about surveys and survey construction than what has been discussed here; this guide should only be used as a starting point. At the very least, researchers using this guide should also consult one or more of the excellent handbooks available on survey research. Finally, researchers should

keep research notes about the procedures used for establishing validity and reliability. Such notes are important for informing the argument that a researcher will need when writing for research publication.

Acknowledgements

Not applicable.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

William W. Cobern  <https://orcid.org/0000-0002-0219-203X>

Betty AJ Adams  <http://orcid.org/0000-0002-8554-8002>

6. REFERENCES

- American Education Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Cobern, W. W. (2000). *Everyday thoughts about nature: An interpretive study of 16 ninth graders' conceptualizations of nature*. Dordrecht, Netherlands: Kluwer Academic Publishers.
- Cobern, W. W., & Adams, B. A. (2020). When interviewing: how many is enough? *International Journal of Assessment Tools in Education*, 7(1), 73-79. <https://doi.org/10.21449/ijate.693217>
- Cobern, W. W., Adams, B. A. J., Pleasants, B. A.-S., Bentley, A., & Kagumba, R. E. (2019, March 31-April 3, 2019). Investigating the potential for unanticipated consequences of teaching the tentative nature of science. Paper presented at the National Association for Research in Science Teaching, Baltimore, MD.
- Cobern, W. W., Gibson, A. T., & Underwood, S. A. (1999). Conceptualizations of nature: An interpretive study of 16 ninth graders' everyday thinking. *Journal of Research in Science Teaching*, 36(5), 541-564.
- Cobern, W. W., Schuster, D. G., Adams, B., Skjold, B., Mugaloglu, E. Z., Bentz, A., & Sparks, K. (2014). Pedagogy of Science Teaching Tests: Formative Assessments of Science Teaching Orientations. *International Journal of Science Education*, 36(13), 2265-2288. Retrieved from <http://bit.ly/RE95xZ>
- Baker, F. B. (2001). *The basics of item response theory*: ERIC Clearinghouse on Assessment and Evaluation. Retrieved from <http://echo.edres.org:8080/irt/baker/final.pdf>
- Bhat, A. (2019). Top 7 Demographic survey questions for questionnaire. Retrieved from <https://www.questionpro.com/blog/demographic-survey-questions/>
- Blair, J., & Conrad, F. G. (2011). Sample size for cognitive interview pretesting. *Public Opinion Quarterly*, 75(4), 636-658. <https://doi.org/10.1093/poq/nfr035>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best Practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149-149. <https://doi.org/10.3389/fpubh.2018.00149>

- Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education*, 15(4), rm4. <https://doi.org/10.1187/cbe.16-04-0148>
- Bruck, A. D., & Towns, M. (2013). Development, implementation, and analysis of a national survey of faculty goals for undergraduate chemistry laboratory. *Journal of Chemical Education*, 90(6), 685-693. <https://doi.org/10.1021/ed300371n>
- Bruck, L. B., Towns, M., & Bretz, S. L. (2010). Faculty perspectives of undergraduate chemistry laboratory: goals and obstacles to success. *Journal of Chemical Education*, 87(12), 1416-1424. <https://doi.org/10.1021/ed900002d>
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: the tailored design method* (4th Ed). Hoboken, NJ: John Wiley & Sons, Inc.
- Fryrear, A. (2016). How to write better demographic survey questions. Retrieved from <https://www.surveygizmo.com/resources/blog/how-to-write-better-demographic-questions/>
- Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: reflections from over 20 years of research. *Educational Psychologist*, 50(1), 14-30. <https://doi.org/10.1080/00461520.2014.989230>
- Haryani, E., Cobern, W. W., & Pleasants, B. A.-S. (2019). Indonesia Vocational High School Science Teachers' Priority Regarding 21st Century Learning Skills in Their Science Classrooms. *Journal of Research in Science Mathematics and Technology Education*, 2(2), 105-133.
- Krosnick, J. A. (2018). Improving question design to maximize reliability and validity. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 95-101). New York: Palgrave Macmillan.
- Lamb, R. L., Annetta, L., Meldrum, J., & Vallett, D. (2012). Measuring science interest: RASCH validation of the science interest survey. *International Journal of Science and Mathematics Education*, 10(3), 643-668. <https://doi.org/10.1007/s10763-011-9314-z>
- Luo, T., Wang, J., Liu, X., & Zhou, J. (2019). Development and application of a scale to measure students' STEM continuing motivation. *International Journal of Science Education*, 41(14), 1885-1904. <https://doi.org/10.1080/09500693.2019.1647472>
- Multon, K. D. (2010). Test-retest reliability. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1495-1498). Thousand Oaks, California: SAGE Publications, Inc.
- Nyutu, E. N., Cobern, W. W., & Pleasants, B. A.-S. (2020). Development of an instrument to assess students' perceptions of their undergraduate laboratory environment. *The Journal for Research and Practice in College Teaching*, 5(1), 1-18. Retrieved from <https://journals.uc.edu/index.php/jrpct/article/view/1492>
- Redline, C. (2013). Clarifying categorical concepts in a web survey. *Public Opinion Quarterly*, 77(S1), 89-105. <https://doi.org/10.1093/poq/nfs067>
- Rosenberg, S. (2017). Respectful collection of demographic data. Retrieved from <https://medium.com/@anna.sarai.rosenberg/respectful-collection-of-demographic-data-56de9fcb80e2>
- Ruel, E. E., Wagner III, W. E., & Gillespie, B. J. (2016). *The practice of survey research: theory and applications*. Los Angeles, CA: SAGE.
- Seidman, I. E. (2006). *Interviewing as qualitative research: a guide for researchers in education and the social sciences, 3rd Edition*. New York: Teachers College, Columbia University.
- Smyth, J. D. (2016). Chapter 16: Designing Questions and Questionnaires. In C. Wolf, D. Joye, T. W. Smith, & Y.-c. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 218-235). <https://doi.org/10.4135/9781473957893>

- Staus, N. L., Lesseig, K., Lamb, R., Falk, J., & Dierking, L. (2019). Validation of a measure of STEM interest for adolescents. *International Journal of Science and Mathematics Education*. <https://doi.org/10.1007/s10763-019-09970-7>
- Taber, K. S. (2018). The use of Cronbach's Alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273-1296. Retrieved from <https://doi.org/10.1007/s11165-016-9602-2>
- Vannette, D. L., & Krosnick, J. A. (2018). *The Palgrave handbook of survey research*. New York: Palgrave Macmillan.
- Willis, G. B. (2016). Chapter 24: Questionnaire pretesting. In C. Wolf, D. Joye, T. W. Smith, & Y.-c. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 359-380). <https://doi.org/10.4135/9781473957893>
- Willis, G. B. (2018). Cognitive interviewing in survey design: State of the science and future directions. In D. L. Vannette & J. A. Krosnick (Eds.), *The Palgrave handbook of survey research* (pp. 103-107). New York: Palgrave Macmillan.