INTERNATIONAL JOURNAL OF MULTIPLE RESEARCH APPROACHES, 2015
VOL. 9, NO. 1, 24–39
http://dx.doi.org/10.1080/18340806.2015.1076757

Routledge
Taylor & Francis Group

# Examining the consequential validity of standardized examinations via public perceptions: a review of mixed methods survey design considerations†

John H. Hitchcock[a], Anthony J. Onwuegbuzie[b] and Heba B. Khoshaim[c]

[a]Instructional Systems Technology Department, Indiana University, Bloomington, IN, USA; [b]Department of Educational Leadership, Sam Houston State University, Huntsville, TX, USA; [c]Department of General Sciences, Prince Sultan University, Riyadh, Saudi Arabia

ABSTRACT

The use of standardized examinations to inform high-stakes decisions is a process that is followed in many countries, and there is merit to understanding the impact that such testing has on any given society. For this reason, some researchers have studied the societal impact of testing to understand the social validity and consequences of assessment. Information gleaned from such efforts can be used to inform testing practices as well as communication efforts with stakeholders. Assessing social validity, and aspects of the consequences of measurement, can be handled in part via survey studies. Because the use of testing is becoming more widespread, and across several nations, anyone interested in such inquiry will need to be able to develop culturally relevant surveys. For this reason, this methodological article describes an exploratory and sequential mixed methods research approach for survey development.

ARTICLE HISTORY
Received 27 September 2014
Accepted 27 June 2015

KEYWORDS
Standardized testing; mixed methods; measurement; culturally relevant surveys; social validity; consequential validity

The use of standardized examinations to inform high-stakes decisions (e.g. college entrance, high school graduation, and grade promotion) is a widespread process that is followed in many countries (Foster, 2010). Such tests are used under the belief that measurement will facilitate decision-making (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 1999, hereafter called The Joint Standards, 1999). Despite the inherent expectation that such assessment is conducted for the benefit of a society, standardized measurement has endured considerable controversy. Consider that, in the USA, the phrase *high-stakes testing* has been a part of the common vernacular for years (see Heubert & Hauser, 1999; The Joint Standards, 1999; U.S. Department of Education, 2000), and this represents a loaded term. It appears that some proponents of standardized measurement embrace the *high-stakes* phrase so as to highlight the importance of education, whereas others use it in a more pejorative manner because important decisions are based on or heavily influenced by a test score.

CONTACT John H. Hitchcock jhhitchc@indiana.edu

†This article is based in part on a previously presented conference paper by Hitchcock, Onwuegbuzie, & Koshaim (2012).

Although standardized assessment is ensconced in all levels of U.S. education, from pre- to graduate school, these measurement systems and how data are used continuously are criticized by national teacher unions, civil rights groups, and any number of grassroots organizations. By way of example, the National Association for the Advancement of Colored People (NAACP) has issued calls for colleges and universities to de-emphasize the use of the American College Test and Scholastic Aptitude Test when making decisions about college admissions, citing them as being unfair and discriminatory (Blair, 1999). In 2012, the NAACP's Legal Defense Fund (NAACPLDF), along with other advocacy groups, filed a federal civil rights complaint concerning the use of standardized measures to make decisions about entry into elite high schools in New York City (NAACPLDF, n.d.). Complicating matters further, the academic measurement community is far from unified on the matter of whether the use of high-stakes tests can act as a fulcrum to improve education (cf. Amit & Fried, 2002; Amrein & Berliner, 2002, Brewer, Knoeppel, & Clark Lindle, 2014; de Lange, 2007; Linn, 2000; Linn, & Gronlund, 2000; Raymond & Hanushek, 2003).

Given the prominence of the argument, it is not surprising that polling organizations routinely query the U.S. citizenry about the role that standardized testing should have in education, and when making high-stakes decisions. After all, politicians and policy-makers have influence over education policy and it behooves them to understand their constituents' views, especially when they rely on measurement results to gauge performance of schools and to make choices about resource allocation that will impact said schools. Hamilton, Stecher, and Klein (2002) summarized the results of several state-level and national polling efforts[1] that date back to the 1990s and concluded that a majority of respondents generally support the use of standardized measurement for high-stakes decisions, although the public believed that such tests have limitations. Polling of this sort likely will continue to be undertaken for the foreseeable future. Indeed, a current U.S. controversy deals with whether student performance on standardized tests should be used in teacher evaluations, and the 2012 annual Phi Delta Kappa/Gallup Poll that assesses attitudes towards schools revealed that the public's attitudes are approximately equally split on the matter (Bushaw & Lopez, 2012). In sum, views of standardized tests are far from unified in the United States of America.

This appears to be the case in other countries as well (Rotberg, 2006). For example, Saudi Arabia has initiated widespread college entrance examination testing (National Center for Assessment in Higher Education, n.d.), and there has been some public concern over the implementation of the program and resultant use of data (Al-Dhakailallah, 2012). Hamilton et al. (2002) raise the point that part of what feeds controversy, at least within the United States of America, is that different stakeholder groups have competing agendas in education. But such differences are exacerbated by the fact that highly technical aspects of testing often are poorly understood within the general public. Another complication arises from test security, which can promote a sense of secrecy. Finally, policy-makers and practitioners often can use testing results in ways that were not intended by test developers and this can undermine valid use of measurement results, which feeds back into public mistrust.

Understanding public perception of testing can have value not only in terms of identifying aspects of a testing procedure that trouble public stakeholders, the process also may yield insights into the psychometric notions of consequential and social validity.

There is considerable scholarship on the notion of consequential validity that have been authored by Education Testing Service personnel (e.g. Messick, 1989, 1994, 1995), authors of measurement texts (e.g. Linn & Gronlound, 2000), and the Joint Standards (1999). Briefly, these authors note that the validity evidence for a given measure is informed in part by the intended and *unintended* consequences of measurement. If negative consequences occur in high-stakes settings, test developers, and policy-makers should investigate and correct the source of the problem. Although discussion of such validity is in the technical literature, it is this aspect of validity, the consequences of testing, that is very much at work in the public realm, and related investigation may offer hints as to what unintended consequences arises when introducing a measurement system in a society.

Social validity deals with the degree to which stakeholders buy into a process or change and the degree to which change is implemented as intended (Nastasi & Hitchcock, 2008). Understanding social validity should be of key interest because stakeholders who do not value or agree with an intervention are likely to undermine its implementation in some form or fashion. Assessing such matters also is consistent with social justice concerns (Onwuegbuzie & Frels, 2013). As a result, it becomes important to assess social validity and make necessary alterations to negotiable program features and/or messages about the program if buy in is considered to be problematic. In the short term, such efforts might inform issues such as cheating, curricular change, and how students study. In the long term, more distal issues arise such as whether the public generally supports high-stakes testing programs. If public perception problems are found, policy-makers and supporting organizations might be able proactively to change procedures that are negotiable, change public messages about testing procedures, engage in education functions, and possibly even use survey results as fodder for different psychometric investigations (e.g. subgroups that have taken strong umbrage with the tests might inform differential item functioning). Overall, proactive assessment of public perception and subsequent use of data could minimize problems.

## Purpose

Lessons learned in the United States of America about consequential and social validity suggest that there can be value in understanding public perceptions of the standardized testing movement. Those who wish to continue with or adopt the use of such examinations might, therefore, consider conducting systematic investigations of societal perceptions of testing to examine the attitudes that different stakeholder groups (e.g. students, parents, and teachers) have about the process. General questions could include: To what extent do different stakeholder groups see standardized testing (say in college entrance examinations) as a benefit or hindrance? To what extent do they value tests? To what extent is there a sense that testing in this process promotes equity and good decision-making? What aspects of the testing process might be improved and what practices should continue without change? Answers to such questions can help policy-makers and testing professionals assess the consequences of their programs and make choices about how to describe the purpose of a test to the wider public. However, a general barrier to gathering high-quality information about public perceptions is that different cultures likely will carry different concerns and values, making the development of an opinion survey a challenging prospect. Therefore, the purpose of this methodological article is to outline a series of mixed methods survey design considerations that can facilitate effective inquiry into how members of a given country (or subpopulations within a country) perceive standardized testing.

In other words, this article does not culminate in the development of a survey, nor do the authors advocate for the adoption of testing programs per se. Rather, the article describes a series of design considerations that readers should consider for developing their own surveys that can help one to understand public perception (either before testing is adopted or after to help understand the consequences of the system). A key point behind the mixed methods approach is that it should yield a survey that is meaningful to target populations of interest.[2]

## Mixed methods research considerations

The clear methodology for understanding public perceptions is the use of surveys. Developing high-quality surveys requires careful attention to item writing, piloting, sampling, and survey dissemination practices. If surveys are to be generated in different countries, survey development could fundamentally be seen as a mixed methods research endeavor, whereby multiple approaches are needed to develop a synergistic approach for developing highly targeted surveys that can work within a cultural and contextual context.

As noted earlier, there is plenty of experience with the use of polls in the United States of America, but there also are considerable cultural and contextual factors to consider if planning to conduct public opinion polls in different countries. For this reason, we outline an exploratory and sequential research approach (cf. Creswell & Plano Clark, 2011; Hitchcock et al., 2005; Onwuegbuzie, Bustamante, & Nelson, 2010) that is predicated on developing an initial understanding of what topics to cover via survey efforts, and how to craft items by systematically interviewing different stakeholder groups and analyzing documents that describe assessment practices. Once the process is complete, this information can be used to inform item crafting and survey dissemination practices that can work effectively in different settings in the country (e.g. trade-offs between different survey modes such as paper vs. web-based for different stakeholder groups).

Take one particular form of a high-stakes test: the college admission examination. As noted previously, there are concerns among parents and educators about how tests are used to make decisions, and how their content might influence the way in which students are being educated in schools. Changes in college admissions might raise questions that are recognizable to those who have observed high-stakes assessment issues in the United States of America. These include:

- To what degree are the new assessments perceived by the public to be valid indicators of cumulative achievement?
- What type of achievement should they represent within subject domains (e.g. memorization as opposed to critical thinking skills?)
- To what extent does aggregate performance on the new assessments suggest a need to change current curricula and teaching styles?
- To what degree do the tests promote fairness in the college admissions process?
- Are any subgroups differentially performing on the tests in ways that might make it easier or more difficult to be admitted to higher education institutions?

When adopting a lens of consequential validity, it becomes important to understand public perceptions of these issues. After all, poor reporting practices might reinforce perceptions that a particular subgroup does not produce students who represent college material, high reliance on testing might compel educators to change curricula, and so on. Further, some groups might have limited trust in the testing process and how information is being used. Hence, conducting such surveys might have value, but as researchers in different countries consider the public surveying process, a number of difficulties can arise. Chief among these is that survey writers will want some assurance that their items adequately cover the domains of interest, or more simply put, there should be evidence that the correct questions are being asked. This might take considerable contextual knowledge, which may be achieved via qualitative inquiry. From there, the specific ways in which items are grouped and worded warrant careful consideration. Finally, there will often be a need to assess whether survey results adequately can inform population inferences, meaning that statistical analyses will be warranted. Given these intersections, survey development in different settings can be bolstered by adopting mixed methods research approaches.

## *The exploratory sequential research design*

Mixed methods research entails the combination of quantitative and qualitative research (see Tashakkori & Teddlie, 1998, 2003, 2010). The point of such a combination is to draw from the strengths of each tradition of research so as to minimize inherent weaknesses associated with a monomethod approach (Johnson & Onwuegbuzie, 2004). The advantage of such mixing is easily demonstrated in survey research. A general model that can be used is the exploratory sequential research design (Creswell & Plano Clark, 2011). The overall approach is straightforward. The phrases "exploratory" and "sequential" denote a design that begins with the use of qualitative research methods to explore phenomena and this is followed up by later quantitative analyses. Before Creswell and Plano Clark's (2011) characterization, ethnographers had used surveys to study culture for the simple reason that it is unrealistic to conduct interviews with large numbers of people (Schensul, Schensul, & LeCompte, 1999). An example of an ethnographic survey can be seen in the aforementioned investigation where perceptions of Sri Lankan adolescents were initially studied using a number of qualitative research methods (i.e. focus groups with students and teachers, interviews with school leaders, observations of classrooms and school processes, and archival analyses), and this led to survey item development. Surveys were, in turn, disseminated to large numbers of students to ascertain whether qualitative research findings held over large groups of respondents (Hitchcock & Nastasi, 2011; Hitchcock et al., 2005, 2006; Nastasi, Hitchcock, Burkholder, Sarkar, & Varjas, 2007). This body of work is germane to potential survey development in other countries because an early assumption in the Sri Lanka investigation was that not enough was initially known about the culture and context to proceed with survey development. As an aside, this was not just because U.S. researchers were involved in the effort. Sri Lankan professors also were involved, but also they did not have the information needed to be certain what questions should be asked or how to word items. This also does not betray any lack of expertise on their part but rather a rigorous accounting of what was known and unknown at the time. This sense of limited knowledge and taking systematic attempts to learn about respondents before dissemination is consistent with empirically driven advice by survey experts. A review of texts by Dillman, Smyth, and Christian (2009), Fowler (2009), and Groves et al. (2009) reveals repeated arguments for a need to understand deeply the populations to be surveyed. A general summary is that strong contextual knowledge is needed for everything from item crafting, to making decisions about survey dissemination, to identifying strategies for improving response rates, to interpreting statistical analyses, and so forth.

Qualitative research approaches are recommended as a way to obtain such knowledge because, for the most part, they entail naturalistic inquiry, emergent design principals, and a focus on how respondents construe their realities and are fundamentally interpretive (Brantlinger, Jiminez, Klingner, Pugach, & Richardson, 2005; Creswell, 2009; Johnson & Christensen, 2012; Patton, 2002; Shank, 2002). Note that this is not a complete list of characteristics but should be sufficient to set the stage for the points made below. Naturalistic inquiry means that qualitative research designs generally attempt to study phenomenon in typical settings, without contriving or manipulating circumstances (there are exceptions). Emergent design refers to the idea that qualitative researchers adapt questions and methods as their understanding improves. There is also an interest in the idea that respondents have a hand in constructing (construing) their own psychological realities. Learning how people make sense of their realities can promote fine-grained understanding that might be a critical outcome when surveying public perceptions of testing; after all, some members of a population might, for example, endorse high-stakes testing for some interventions (e.g. allocation of significant resources to struggling students while promoting them to the next grade) but not for others (allowing grade retention, that is having a student repeat a grade, as a fulcrum for allocating such resources).

With this in mind, most qualitative inquiry involves the use of some combination of interviews, focus groups, observations, and archival analyses. There are several variants to these different techniques and it is beyond the scope of this article to review all of the choices. Instead, consider the likely advantages of conducting a series of interviews and focus groups with different representatives of likely subgroups. The general focus of this data collection phase should be to consider what concerns, if any, are held about college admissions tests. Interviewers can explore whether there are concerns with a specific test or subtest. Questions can be asked about assessment practices, security procedures, use of data, perceived benefits of testing, and so on. This general line of inquiry can be undertaken with individuals or in focus groups. These data collection techniques are quite different and used for different reasons. Interviews do, of course, generally allow for deeper exploration because time is spent with a single participant. For this reason, interviews are preferable when gathering data from people with expertise dealing with complex topics (e.g. testing professionals who can provide information about test construction, dissemination procedures, security measures, policy-makers who might have information about legal and political plans) or also when dealing with private behavior (e.g. reasons for why test takers were compelled to cheat on a test). Focus groups by contrast can be more efficient than are interviews because they are conducted with groups of approximately six to 12 people at a time, but this also generally means that fewer questions can be asked. One advantage to focus groups is that they offer an opportunity to observe social processes such as whether consensus on a controversial topic is reached over time. Onwuegbuzie, Dickinson, Leech, and Zoran (2010) provide a current literature review of focus group practices and frameworks for assessing group consensus and nonverbal behavior, as well as how to take a mixed methods research approach whereby procedures such as micro-interlocutor analyses (e.g. how many respondents reply to a question, the length of their responses, number of dissenters or a particular view) are conducted. These techniques may be particularly relevant to understanding different stakeholder perceptions of tests and testing procedures because opinions often are reflective of a social process. To promote naturalistic inquiry, there can be advantages to conducting focus groups in settings where respondents are used to attending (e.g. home schools). Now consider that these steps can be repeated with several different stakeholder groups. Obvious choices would be students, parents, teachers, school administrators, testing professionals, and policy-makers. More fine-grained information might be gained by further splitting these demarcations into subgroups, such as students who have taken given tests versus those who are preparing. Of course, doing this work in different parts of a country would be useful if the intent is to develop national surveys. Additional details may be gained by systematically reviewing archival information that might include newspaper and Internet accounts of testing procedures, testing materials, promotional information, policy documents, and so on. Finally, further understanding could be gleaned from observing relevant practices like test administration and efforts to prepare students for tests.

Qualitative data analysis is not a trivial step, and there are an enormous number of sources that offer general analytic techniques. Likewise, there are a number of general techniques that describe different procedures for how to promote the rigor and credibility of this sort of work (e.g. Bratlinger et al., 2005; Creswell, 2009; Denzin & Lincoln, 2005; Lincoln & Guba, 1985; Nastasi & Schensul, 2005; Onwuegbuzie & Leech, 2007; Patton, 2002; Shank, 2002). Like other general discussions of qualitative research methods, it is beyond the scope of this article to go into any detail. However, we mention the issue here so that readers can be alerted to the existence of credibility techniques should this work ever be pursued.

### *Item writing*

Survey development can begin with specifying the purpose and structure of an instrument, which, in this case, can be more effectively undertaken after the qualitative phase is complete and there is a sense of the merits and concerns held by the public with respect to testing. Once researchers obtain a good sense of what questions to ask, wording the questions in an effective way and considering other issues like the order in which to place them could have enormous influence over how respondents answer questions (Dillman et al., 2009). To apply this idea in the survey world, consider two concerns offered by Dillman et al. (2009): wording and anchoring, and a related concern, which is back-translation. To start with a simple example, phrases that are needlessly complex can yield confusion. Consider the differences between the items: (a) should colleges not rely exclusively on admissions tests when making application decisions? and (b) should colleges rely exclusively on admissions tests when making application decisions? The first choice can confuse respondents, especially when using a common response scale indicating level of agreement.

Parsing semantics also is an issue. By way of example, one of the co-authors of this article previously developed a short survey to assess the types of professional development that teachers have completed during the prior year of their careers. During pilot work, it was discovered that when teachers were asked about the number of hours of professional development that they received over the year, responses ranged from zero to thousands of hours. The problem was that one teacher viewed her teaching experience to be a form of professional development, so she estimated how many hours she worked. Another teacher considered only time spent in college classrooms as professional development and so did not consider workshops. In order to obtain meaningful data, the related item had to be reworded. Clearly, wording and vocabulary become critical factors in any survey effort. This is a straightforward point but now consider how much more complex wording choices can be when dealing with back-translation (i.e. when a survey is developed in one language, translated to another language, and then translated back to the original language). Unique cultural factors can interact with survey design in ways that have an influence on any data that are collected. We contend that it can be highly difficult to understand such influences without the benefit of qualitative data collection and analyses. A variant of these procedures long has been recognized in the survey literature, where experts recommend engaging in cognitive interviews, which are also sometimes called *cognitive labs* (Willis, 2005).

For yet another example of how culture can influence survey design, a common attitudinal assessment scheme is to use a Likert-format response option or one of its variants (e.g. asking respondents to rate whether they strongly agree with a statement, agree, disagree, and so on). Hitchcock et al. (2005) developed a survey scheme to understand culturally specific values and competencies of Sri Lankan students as part of a wider research project. During initial pilot work, the principal investigator of the project (Nastasi) found that the standard 5-point response scale led to some confusion among Sri Lankans. This was in part because surveys were not commonly used in that country and, in part, because of confusion among the populace around the difference between strongly agreeing and just agreeing with a statement. For these reasons, a simpler 3-point response scale (agree, neutral, and disagree) was adopted.

The central point to all of this is to establish the idea that item crafting is subject to any number of cultural and contextual factors. For this reason, simple translation of existing surveys about attitudes of high-stakes testing likely will yield inferior measurement. In short, a different approach is recommended for anyone who wishes to assess public attitudes towards college admission testing, and later, other forms of high-stakes assessment.

Recall that mixed methods research approaches use the strengths of one technique to compensate for the weaknesses of another. The strength of qualitative research techniques would be to explore what different stakeholder groups think about tests and testing procedures. The approaches can yield deep insights into any number of topics related to the utility of tests, their consequential validity, the procedures followed, and so on. But the shortcoming to interviews is that the capacity to generalize what is learned tends to be limited. Policy-makers should be interested in obtaining perceptions from representative samples of different stakeholder groups. Surveying public perceptions is, after all, a national endeavor that will be influenced by national opinion. So, the transition here would be to generate survey items and coordinate these items in ways that are informed by prior qualitative work. In-depth qualitative exploration will provide fodder for what questions to ask in order to gauge perceptions of particular stakeholder groups about college admissions tests, and how to go about asking them. The point here would be to avoid problems, like the ones noted earlier, that can arise from not having a good grasp of terminology and contemporary perceptions and issues.

### *Piloting*

Once any particular survey is drafted, it should be subjected to a few levels of piloting. Willis (2005) and Fowler (2009) outline steps where surveys can be assessed by, in essence, interviewing respondents as they fill out initial versions. Typical points of consideration include, but are not limited to: (a) whether items seem adequate to assess the intended topic; (b) whether any words/items are problematic in that they are confusing, offensive, or evince some level of ignorance that will undermine trust; (c) whether the order of items appears to be problematic in any way; (d) whether directions are clear and accurate and timing estimates are accurate; (e) whether any technological applications (e.g. web-based surveys) are apparent; (f) whether the visual layout is easy to follow and properly sub-divides items into logical sections; and (g) whether response options match up well with items; and so on. During this testing phase, survey developers should observe the respondent as he/she completes the instrument. In addition, respondents should be asked to think out loud as they react to and complete items. As a general rule, this process should be repeated until no problems are detected across a few cycles. Once this stage is complete, it becomes critical to pilot test the survey with a proxy sample. Again, respondents should be given an opportunity to provide feedback but, at this stage, a sample should be collected so as to obtain initial evidence on the internal consistency of scores yielded by scales, which would entail obtaining Cronbach alpha estimates. Johanson and Brooks (2010) completed some simulation studies in sample size requirements for yielding stable estimates. Groups of approximately 30 are generally going to be adequate.

### *Sampling considerations*

Dillman et al. (2009) distinguish surveys from survey design; the latter deals more with dissemination choices. Clearly, some variant of random sampling would be ideal because the hope is to be able to engage in probabilistic generalization of sample observations back to populations of interest. Groves et al. (2009) offer some good overviews of different design options although there are many sources on this topic. A key consideration, of course, is identifying a sample frame (this is the formal list from which a random sample of participants will be drawn) and assessing any coverage error (Dillman et al., 2009), meaning the degree to which the frame does not adequately capture the target population of interest. An alternative approach may be to use mixed methods sampling designs that might entail a combination of random and purposeful sampling at different stages within an overall research plan (Onwuegbuzie & Collins, 2007). Indeed, Onwuegbuzie and Collins (2007) provide details on sequencing techniques, sample sizes that can be considered for different facets of a design, and how different approaches might be mixed (or crossed) to develop an in-depth sense of sample representativeness. One might, of course, use purposeful selection techniques to promote the likelihood that well-informed stakeholders are interviewed, or join focus groups, and random sampling techniques can be applied once a survey is finalized and there is a desire to make population-based inferences.

Once a sample is selected, the next considerations are working through what survey modalities and incentives might be applied to minimize nonresponse rates. Here, prior qualitative research work can, again, yield useful information. It is reasonable to assume that web-based survey delivery will work well with students who take college admissions tests, but this assumption becomes increasingly tenable as the focus shifts to students who do not take the tests, parents, the elderly, and those working in any geographic region where Internet and/or personal computer access is problematic. In these cases, using different modes such as the mail, automated phone calls, and so on might be necessary. But a key point is that prior questions that elicit advice on how to deliver surveys might turn out to be useful.

### *Response rates and incentives*

Another topic related to design deals with response incentives. In the United States of America, there is considerable information that can inform decisions about response rates that is based on Social Exchange Theory (Dillman et al., 2009). The general premise of the theory is that people are more willing to enter exchanges when the perceived benefit of doing so outweighs the perceived cost. For this reason, considerable thought should be put into both issues of reward (e.g. incentives, how responding can be beneficial not only to the respondent but also to groups that are important to the respondent) but also cost (e.g. time to complete the survey, inconveniences caused by furnishing responses, any distress the process might cause). As it turns out, much of what informs this process is how information is perceived. In the United States of America, it has generally been found that it is best not to present incentives to complete surveys in a quid pro quo manner because this promotes a sense of economic exchange, and most incentives will not begin to reach respondents' price points. So, rather, it is best to present incentives as small tokens of appreciation for the respondent giving up valuable time. There is also empirical evidence that directions should be carefully worded so as to avoid subordinating or otherwise belittling potential respondents, because doing so increases the perception of cost, which, in turn, decreases the chances that people will complete the survey. Small wording choices like "you must do this" versus "please do this" can have an important influence. This is seemingly obvious but the literature suggests that survey writers often demonstrate a poor sense of Social Exchange Theory to their detriment (Dillman et al., 2009). Furthermore, drivers of what will lead people to perceive something as a cost or benefit are likely to be highly culturally embedded. In many countries, there are any number of micro-cultures that should give survey writers some pause. One more point to make here is that Social Exchange Theory predicts that, as topics and questions are perceived to be more salient and interesting, the likelihood of responding will increase because the benefit of doing so increases. This gets back to the benefits of knowing what questions to ask, how to word them, and where to place them in an instrument (attention-grabbing items should generally be placed toward the beginning of a survey so as to get a person to start responding). To study these matters when designing any particular survey, refer back to the qualitative research stage of the sequential exploratory design.

### *Mixed methods research and survey error*

Consider the four sources of survey error described by Dillman et al. (2009) and Groves et al. (2009): coverage, sampling, nonresponse, and measurement error. Coverage error considers whether sampling frames (i.e. a list) adequately represent the population of interest. For example, if policy-makers wish to survey students who have yet to take college admissions tests, it would be necessary to work out a frame from which to sample potential respondents. And the degree to which the frame fails to capture all students is the degree to which there are coverage error problems. Sampling error is simply a matter of sample size, wherein there is an inverse relationship between sample size and statistical error. Unless researchers are dealing with difficult-to-find subgroups, it is likely that samples will be more than adequate for yielding tolerable confidence intervals around parameter estimates. Nonresponse error can, however, be a serious concern because this can generate bias in estimates. Modern methods in handling missing data such as multiple imputation and maximum likelihood procedures (Enders, 2010) can provide important statistical corrections, but a general maxim is that the best way to deal with missing data is to avoid them (Allison, 2002). Here, prior qualitative inquiry with an eye toward Social Exchange Theory can be critical, as noted earlier. But when survey results are undermined by missing data, qualitative research methods can, again, be potentially useful. Briefly, the statistical literature has identified mechanisms of missing data: missing completely at random, missing at random (MAR), and missing not at random. For purposes of statistical inference, the first mechanism is the most desirable because this essentially means that the collected data are simply a random subsample from the desired sample. The only penalty for the missing data is decreased statistical power (or in the parlance of survey work, increased sampling error). When data are MAR, this suggests that the probability of missing data is not related to a dependent variable itself but might be related to other variables (e.g. demographic variables). As a simple example, respondents might have failed to report their personal income but this is not because of how much they make; rather, perhaps the age of the respondents appears to predict the likelihood that this information is furnished. The most pernicious missing datum is the last one, and is assumed when missingness on a variable appears to be related to the variable itself (e.g. how much a person makes per year appears to drive the likelihood that they refuse to divulge this information). Corrective procedures are thought to be reasonable when data are MAR, but one cannot be sure that this assumption is accurate. Qualitative research methods might be of use when assessing the missing data mechanism because they might learn through interviews why respondents refused to answer some items. Usually one engages in exploratory regression to diagnose the missing data mechanism, but we suggest that exploratory interviews can be used in several ways to examine the concern. First, the above *think-aloud* procedures might alert researchers to items that might be viewed as invasive, written in a harsh manner, or overly personal. It might also be the case that follow-up efforts can be pursued with subgroups of people to learn about why responses on given items were not offered. Note that such exploration would not necessarily entail in-depth interviewing, but rapid reconnaissance procedures to determine whether one mechanism is more reasonable than another to invoke.

Finally, there is measurement error, a topic well familiar to anyone with minimal psychometric training. Failure to ask appropriate questions and problems with wording and ordering them once they are identified can yield serious concerns with this form of error in survey work (recall the aforementioned considerations with back translating). This article is predicated on the assumption that prior qualitative inquiry can be useful when developing items to address attitudes around college admissions surveys, thereby reducing measurement error.

### *Additional statistical analysis ideas*

This section provides a brief overview of general analytic directions that can be pursued, depending on whatever research questions that might arise. As was the case when introducing qualitative research procedures, details are not provided. The purpose here is not to offer a primer on common statistical procedures, but rather to offer general analytic ideas that can be used within the exploratory sequential research design. In this stage, researchers will have sequenced from qualitative inquiry and used some sampling scheme to obtain enough responses to engage in parametric statistical analyses. Once estimates are derived, it will often be possible to consult qualitative findings to aid in interpretation and to guide subsequent research questions. Before proceeding, one point that is not to be lost is that, clearly, descriptive analyses of data should yield sufficient actionable information that can inform policy development about testing procedures, efforts to educate the public, and so on.

Larger scale efforts can open the door to mixed methods construct validation using exploratory factor analyses. Guadagnoli and Velicer (1988) performed a series of data simulations and found that sample sizes of 300 will generally be able to yield stable coefficients in exploratory factor analysis, and sample sizes that are considerably smaller can be reasonable depending on the size of coefficients and number of items. Once a factor solution is reached, this can be compared to qualitative themes in order to engage in cross-method triangulation as part of construct identification and validation efforts (Hitchcock et al., 2005). Some researchers may wish to engage in confirmatory factor analyses (CFAs) for cross-validation purposes and further to explore constructs. This can be possible with sufficiently large data sets assuming random, independent subsets are drawn for each factor analytic approach. These steps by themselves might yield critical insights on attitudinal constructs when it comes to understanding how the populace perceives college admissions tests and related procedures. In any case, factor analyses in any form can reduce the number of data points, which can simplify further analyses.

As a final note for this section, other advanced techniques might be of great use for understanding data. CFA could, of course, be expanded to structural equation modeling approaches and this might yield any number of useful insights. Keeping in mind that the overall design scheme is based on the idea that item writing and survey development was informed by qualitative inquiry, it can be a novel confirmation exercise to share aggregate findings from models with early study participants (or a new sample) further to query what was learned about stakeholder perceptions. Yet, other techniques, like Item Response Theory approaches, could be used for attitudinal scale development (we assume Rasch modeling should be fine; a 2PL model would probably be unnecessary because there should be no guessing per se when sharing one's opinion).

### *Utility of information*

Results of analytic work can and often should be shared with subgroups of survey participants so as to facilitate decision-making about how effectively to use findings. A central concern to any of the types of surveys described here is to assess public opinions about testing and its processes, as well as to gain insights about how college admissions testing has impacted any given society. A reasonable expectation is some controversy and discord will be discovered from these efforts. When such findings are uncovered, they should inform subsequent action. Suppose test security contributes to a sense of distrust among a large number of respondents. It might be important to engage in messaging and advertising procedures that explain the reasons for security. Imagine that some groups do not perceive college admissions tests to be fair. This could actually provide fodder for differential item function analyses in the name of consequential validity to check on the accuracy of the perception if the particular psychometric details are not already known. Assuming that there is no evidence the test is indeed unfair to some subgroup, it might be in policy-makers' interest to determine how to explain fairness.

A potentially critical aspect of consequential validity could relate to curricular change. If it turns out to be accurate that the aspects of creativity and application of knowledge measured by college entrance examinations is problematic because pedagogy emphasized rote memorization of information, policy-makers might be well positioned to advance their plans for transforming education practice. Survey work could be completed with educators to learn more about their teaching procedures, and results from surveys might bring about critical improvements not only among the current teaching force but also in terms of how future teachers are prepared. This again comes back to qualitative research investigation. Messages might be unheard, misinterpreted, or ignored, especially because some topics are simply difficult to explain to a lay audience. Careful consideration of how much information is too much, or too little, is warranted. For this reason, reviewing strategies for explaining test qualities and procedures, both in terms of how and what to say, could be undertaken in focus groups and after conducting interviews with gatekeepers (i.e. people who have strong localized knowledge and expertise about a particular subgroup that a policy-maker might attempt to reach).

## Conclusions

This article is predicated on the propositions that:

(1) Assessing public perceptions (via surveys) of a testing system can promote understanding of consequential and social validity, and

(2) Contextual and cultural factors are not trivial, and should be expected to influence all facets of survey work, ranging from initial conceptualization to item writing and organization, to making decisions about modality of survey delivery, analyses, and interpretation, and how to use information to affect change.

High-stakes standardized testing is becoming more of the norm across a number of countries. We fundamentally see value in assessment information as a way to facilitate decisions but we also remain sensitive to abuses of resulting data (both intended and unintended). We see wisdom in attending to consequential validity and taking the time to understand public perception of testing. In some cases, such perceptions might motivate more careful crafting of messages on the part of policy-makers and test developers and, just as importantly, signs of mistrust, confusion, and evidence of marginalization should yield careful review and revision of testing procedures. Indeed, addressing consequential validity, in part, by understanding public perceptions of assessment, is consistent with a social justice perspective.

A key difficulty, however, with developing a program of surveys across different countries is that cultural differences, and the context of testing, might render existing surveys useless and, thus, culturally specific instruments should be developed. This process can be informed by mixed methods research designs such as an exploratory-sequential mixed methods design, and we hope that this article provides a sufficient outline for how such a design might be applied. Survey development would be a lot easier if existing items/instruments could simply be adopted via translation and delivered to some random sample of potential respondents. This process can be informed by an exploratory-sequential mixed methods research design. This article outlined the process of how to apply such design and we hope that some of the ideas will be considered.

## Notes

1. Polls often entail short, simple items. In these cases some of the methods described here might be unnecessary. An operating assumption is that readers will be interested in surveying nuanced concerns among the citizenry in a given country.
2. – There is a philosophical orientation behind why the authors do not offer a simple set of polling questions that might be translated from prior surveys used in the USA and other countries. We assume cultural and contextual factors are so pervasive and complex that it is unrealistic to assume that key elements needed to write a survey are currently in place. Namely, it might be premature to know what questions to ask and how to go about asking them in a particular setting. After all, a simple problem with any survey is that failure to ask appropriate questions, in an appropriate manner, will yield data of limited utility. Students might, for example, have some strong opinions about being subjected to college entrance examinations, but we will not discover their views if we do not know what questions to pose.

## References

Al-Dhakailallah, D., (2012, December). *Content analysis of Saudi newspaper articles about NCA*. Paper presented at The First International Conference on Assessment and Evaluation, Riyadh, Saudi Arabia.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

American Education Research Association, American Psychological Association, & National Council on Measurement in Education (1999) *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Amit, M., & Fried, M. N. (2002). High-stakes assessment as a tool for promoting mathematical literacy and the democratization of mathematics education. *Journal of Mathematical Behavior*, *21*, 499–514.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18). Retrieved from http://epaa.asu.edu/epaa/v10n18/

Blair, J. (1999). NAACP criticizes colleges' use of SAT, ACT. *Education Week, 19*(4), p. 10.

Brantlinger, E., Jiminez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children*, *71*, 195–207. doi:10.1177/001440290507100205

Brewer, C., Knoeppel, R. C., & Clark Lindle, J. (2014). Consequential validity of accountability policy: Public understanding of assessments. *Educational Policy*, 1–35. doi:10.1177/0895904813518099

Bushaw, W. J., & Lopez, S. J. (2012). *The 44th annual Phi Delta Kappa/Gallup Poll of the public's attitudes toward the public schools: Public education in the United States: A nation divided*. 94(1), kappanmagazine.org. Retrieved from http://www.pdkintl.org/poll/docs/2012-Gallup-poll-full-report.pdf

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (2005). The discipline and practice of qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.) *The Sage handbook of qualitative research* (3rd ed., pp. 1–32). Thousand Oaks, 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.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Foster, D. F. (2010). Worldwide testing and test security issues. *Ethical Challenges and Solutions*, *20*, 207–228. doi:10.1080/10508421003798943

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

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley.

Guadagnoli, E., & Velicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*, 265–275. doi:10.1037//0033-2909.103.2.265

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). *Making sense of test-based accountability in education.* Washington, DC: Rand.

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: Testing for tracking, promotion and graduation.* Washington, DC: National Academy Press.

Hitchcock, J. H., & Nastasi, B. K. (2011). Mixed methods for construct validation. In W.P. Vogt & M. Williams (Eds.), *The Sage handbook of methodological innovation* (pp. 249–268). Thousand Oaks, CA: Sage.

Hitchcock, J. H., Nastasi, B. K., Dai, D., Newman, J., Jayasena, A., Bernstein-Moore, R., Sarkar, S., & Varjas, K. (2005). Illustrating a mixed-method approach for validating culturally specific constructs. *Journal of School Psychology*, *43*, 259–278. doi:10.1016/j.jsp.2005.04.007

Hitchcock, J. H., Onwuegbuzie, A., & Koshaim, H. (2012, December). *Social validity and college entrance exams in Saudi Arabia.* Paper presented at The First International Conference on Assessment and Evaluation, Riyadh, Saudi Arabia.

Hitchcock, J. H., Sarkar, S., Nastasi, B. K., Burkholder, G., Varjas, K., & Jayasena, A. (2006). Validating culture- and gender-specific constructs: A mixed-method approach to advance assessment procedures in cross-cultural settings. *Journal of Applied School Psychology*, *22*(2), 13–33. doi:10.1300/J370v22n02_02

Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement*, *70*, 394–400. doi:10.1177/0013164409355692

Johnson, R. B., & Christenson, L. (2012). *Educational research: Quantitative, qualitative, and mixed approaches* (4th ed.). Thousand Oaks, CA: Sage.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*(7), 14–26. doi:10.3102/0013189X033007014

de Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester, Jr. (Ed.). *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (Vol. 2, pp. 1111–1142). Charlotte, NC: Information Age.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16. doi:10.3102/0013189X029002004

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Old Tappan, NJ: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13–23. doi:10.2307/1176219

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

NAACP's Legal Defense Fund. (n.d.). *LDF and others file complaint against New York City specialized high schools challenging admissions process*. Retrieved from http://www.naacpldf.org/update/ldf-and-others-file-complaint-against-new-york-city-specialized-high-schools-challenging-admi

Nastasi, B. K., & Hitchcock, J. H. (2008). Evaluating quality and effectiveness of population-based services. In B. J. Doll & J. A. Cummings (Eds.), *Transforming school mental health services: Population-based approaches to promoting the competency and wellness of children* (pp. 245–276). Thousand Oaks, CA: Corwin Press with the National Association of School Psychologists.

Nastasi, B. K., Hitchcock, J. H., Burkholder, G., Sarkar, S., & Varjas, K. (2007). Mixed methods in intervention research: Theory to adaptation. *Journal of Mixed Methods Research*, *1*, 164–182. doi:10.1177/1558689806298181

Nastasi, B. K., & Schensul. S. L. (2005). Contributions of qualitative research to the validity of intervention research. *Journal of School Psychology*, *42*, 177–195. doi:10.1016/j.jsp.2005.04.003

National Center for Assessment in Higher Education. (n.d.). Retrieved October 25, from the Wiki: http://en.wikipedia.org/wiki/National_Center_for_Assessment_in_Higher_Education

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research*, *4*, 56–78. doi:10.1177/1558689809355805

Onwuegbuzie, A. J., & Collins, K. M. T. (2007). A typology of mixed methods sampling designs in social science research. *The Qualitative Report*, *12*, 281–316.

Onwuegbuzie, A. J., Dickinson, W. B., Leech, N. L., & Zoran, A. G. (2010). Toward more rigor in focus group research in stress and coping and beyond: A new mixed research framework for collecting and analyzing focus group data. In G. S. Gates, W. H. Gmelch, & M. Wolverton (Series Eds.) & K. M. T. Collins, A. J. Onwuegbuzie, & Q. G. Jiao (Vol. Eds.), *Toward a broader understanding of stress and coping: Mixed methods approaches* (pp. 243–285). The research on stress and coping in education series (Vol. 5). Charlotte, NC: Information Age Publishing.

Onwuegbuzie, A. J., & Frels, R. K. (2013). Introduction: Towards a new research philosophy for addressing social justice issues: Critical dialectical pluralism 1.0. *International Journal of Multiple Research Approaches*, *7*, 9–26. doi:10.5172/mra.2013.7.1.9

Onwuegbuzie, A. J., & Leech, N. L. (2007). Validity and qualitative research: An oxymoron? *Quality & Quantity*, *41*, 233–249. doi:10.1007/s11135-006-9000-3

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Raymond, M. E., & Hanushek, E. A. (2003). Shopping for evidence against school accountability. In W. J. Fowler, Jr. (Ed.), *Developments in school finance* (pp. 119–130). Washington, DC: National Center for Education Statistics.

Rotberg, I. C. (2006). Assessment around the world. *Educational Leadership*, *64*(3), 58–63.

Schensul, S., Schensul, J., & LeCompte, M. (1999). *Essential ethnographic methods*. Walnut Creek, CA: Altimira Press.

Shank, G. D. (2002). *Qualitative research: A personal skills approach*. Upper Saddle River, NJ: Merrill Prentice Hall.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (2003). *Sage handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (2010). *Sage handbook of mixed methods in social and behavioral research* (2nd ed.). Thousand Oaks, CA: Sage.

U.S. Department of Education. (2000). *The use of tests as part of high-stakes decision-making for students: A resource guide for educators and policy-makers*. Washington, DC: U.S. Department of Education Office for Civil Rights.

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