ANNUAL REVIEWS

Annual Review of Organizational Psychology and Organizational Behavior

Measurement Development and Evaluation

Michael J. Zickar

Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43402, USA; email: mzickar@bgsu.edu

Annu. Rev. Organ. Psychol. Organ. Behav. 2020. 7:213–32

First published as a Review in Advance on October 24, 2019

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

https://doi.org/10.1146/annurev-orgpsych-012119-044957

Keywords

psychological measurement, scale development, factor analysis, item response theory, psychometrics, test theory

Abstract

Psychological measurement is at the heart of organizational research. I review recent practices in the area of measurement development and evaluation, detailing best practice recommendations in both of these areas. Throughout the article, I stress that theory and discovery should guide scale development and that statistical tools, although they play a crucial role, should be chosen to best evaluate the theoretical underpinnings of scales as well as to best promote discovery. I review all stages of scale development and evaluation, ranging from construct specification and item writing, to scale revision. Different statistical frameworks are considered, including classical test theory, exploratory factor analysis, confirmatory factor analysis, and item response theory, and I encourage readers to consider how best to use each of these tools to capitalize on each approach's particular strengths.

INTRODUCTION

Psychological measurement is at the heart of the organizational sciences. Most of the constructs that we include in our studies cannot be directly observed. That is, there is no magic wand that we can wave over an employee to determine her innovative potential or self-esteem. As such, we need to develop the "magic wands," instruments that can provide meaningful results to help us advance our science and practice. In this article, I review some of the best practices related to measure development, review key psychometric frameworks that can guide measure development, and focus on practical decisions that need to be considered. Throughout this article, I make four overall points:

1. Measurement development is not a mathematical process and there is no cookbook of procedures that need to be followed. Instead, there are a set of best practices that should be considered.
2. The process of measurement development is iterative and ongoing. No single study is likely to result in a particular scale being finalized.
3. There are lots of statistical techniques that can be used to help guide the process. Sophisticated techniques are not always better. This article reviews several techniques and explains the place and purpose for each approach.
4. Measurement development at its best is a set of processes that can lead to important conceptual insights and discovery.

In this article, I review research from organizational science researchers, psychometricians, and educational assessment experts. Measurement development is truly interdisciplinary with important advances being made in various disciplines. This review focuses on two stages of measurement development and then measurement evaluation. As I mention throughout, the two stages are not independent but in many ways rely on each other; as such, when developing a new scale, one must anticipate measure evaluation as well as what will guide measurement development.

MEASUREMENT DEVELOPMENT

In this section, I focus on construct explication, the process of detailing the theoretical construct that one's scale aims to measure. Then I focus on practical issues related to item writing, detailing existing research on various item stem and response formats and features. **Table 1** summarizes some of the key steps in the processes outlined in this section.

Construct Explication

This is an area in which organizational researchers can learn a lot from high-stakes educational assessment. Test developers in educational testing often start with a well-defined map of the

Table 1 Measure development steps

Steps	Tasks
Construct explication	1. Develop a working definition and structure of the construct being measured. Are there relevant subfacets that should be anticipated? Outline the specific purposes for the new measure. 2. Research similar constructs and scales to determine whether the new scale is necessary.
Item writing	1. Before writing, decide on the most appropriate format for items. Review item writing best practices. 2. Write significantly more items than needed. 3. Have people who did not write the items review them for clarity and evaluate them for relevance to the construct explication. 4. Edit items as needed and discard items that cannot be fixed.

knowledge base being assessed, that is, content specifications. For example, Fives & DiDonato-Barnes (2013) argue that educational test developers should start educational tests with a table of specification (TOS) that links important content that needs to be tested along with number of items for each content domain plus the desired difficulty, as well as the desired format of the items. They argue that coming up with this TOS before writing the test helps ensure that the test has content validity and helps solidify the writing process. Buckles & Walstad (2008) detail the development of a national assessment of economics education. In this case, where the test has consequences that affect many test takers, a series of content specifications were decided by a committee of experts and communicated to test takers to ensure fairness. It is easy to see the importance of content specifications for high-stakes educational testing where having a set domain of items can provide a sense of fairness to test takers as well as facilitate item writing, which is often done by a team of item writers. For most organizational scales used purely for research, the stakes might be lower (e.g., huge numbers of individuals will not be denied educational opportunities because of the new scale of self-esteem), but the benefits of content specification can be helpful in clarifying the construct(s) and helping guide future validation efforts. For organizational scales used for research, one should construct an operational definition of the construct being assessed and outline in advance a nomological net, specifying how one envisions the scale fitting in with other scales of related constructs. Clark & Watson (2019) highlight the importance of doing a thorough literature review at this stage to make sure the proposed scale is distinct from previous efforts. Possible facets or subdimensions should be planned or anticipated as well. Too often, organizational scales are constructed with little more than an intuitive working definition in one's head, and often the process of item writing helps define the construct. The new scale was created for a particular purpose and that purpose should be explicated in as much detail as possible in advance, and that process of construct explication should help guide item writing.

One must determine in advance what the expectation is for the new scale relating to existing scales of similar construct. This process can help identify how the items need to be written to provide distinct construct measurement. With the problem of construct proliferation (see Shaffer et al. 2016), one must determine in advance whether the scale really needs to be created and, if so, how it will relate to existing measures. The jingle fallacy is that two scales measure the same construct because they share the same construct name, whereas the jangle fallacy is the converse, that two scales with different names measure the different constructs. The jangle fallacy is especially common in the organizational sciences with different scales that share the same name having different components. For example, there were scales that were labeled as extraversion that were correlated modestly with each other. Some scales included a dominance component to extraversion, whereas others focused more on sociability. The key lesson here is that scales are not defined by their names, but by the content of the items. Never trust a scale without examining the individual items. This problem could be avoided by having scale developers spend more time on delineating constructs.

In a new literature, one of the first strategies is for people to develop new scales to measure important constructs in that domain. Several scales aiming at measuring the same construct may be developed in a short amount of time. In addition, as scales for constructs develop, an individual scale may be modified slightly to adapt to theoretical changes and nuances. When proposing a new construct, one must demonstrate that the scales measuring the new construct are conceptually distinct from pre-existing scales.

Item Writing

After a construct has been defined, the next step is to write items. Although item writing has traditionally been thought of as an art, there have been increasing amounts of research into best

item writing practice. In some areas, in fact, item writing has been replaced by item generation done by computer algorithms. Automatic item generation (AIG) has been used in domains that are extremely well-defined, such as arithmetic, in which the features of items can be used to predict the likely difficulty and discrimination of an item. For example, in multiplication, we know that having numbers with several digits increases difficulty, just as in subtraction where the need to carry over increases difficulty. Therefore, an algorithm that uses a random number generator can be used to create items on the fly, thus making a unique test for each individual, along with a scoring mechanism that can account for differences in items to place scores on the same scale. For most of the types of items, however—that organizational scientists are interested in—there just is not the understanding of the relation between item features and content with item functioning. AIG is not a possibility for attitudinal and personality constructs at this point (see Gierl & Haladyna 2013 for more information on AIG).

Many chroniclers describe item writing as an art, which implies that there are no strict rules to guide item creation, and that there are multiple ways to write good items. Although both of those statements are likely true, there is a growing literature on what makes items function well such that they can be used to guide item writing. Haladyna & Rodriguez (2013) provide a comprehensive set of guidelines for writing items for selected-response items (e.g., multiple choice format) based on empirical research. These guidelines should be studied by all who write selected-response items and include guidelines such as "place options in logical or numerical order" and "avoid using the options *none-of-the-above*, *all-of-the-above*, and *I don't know*" [Haladyna & Rodriguez 2013, p. 91 (italics in the original)]. They also include guidelines for survey items, which include attitudes, preferences, opinions, and noncognitive traits. Their guidelines include "use as few words as possible in each item stem and options" and "verbally label all response categories" (Haladyna & Rodriguez 2013). Haladyna (along with many colleagues) has made a career out of studying what makes a good item. Most of Haladyna's work has focused on educational assessment, and so his work is less known by organizational researchers. This is unfortunate. Key takeaways from reading this literature are that items should be written in simple and direct language, avoid slang or colloquialisms, and avoid ambiguous language as much as possible. In short, the result should be variance that is due solely to variations on the construct of interest, not reading ability or special cultural knowledge.

Regardless of the type of scale or construct, all good item writing must start with a well-defined construct. Therefore, the work outlined in the previous section is crucial. Write more items than you will need. Although there are guidelines out there about this, the likely ratio of items needed to be written compared to the final target number likely depends on the expertise of the item writers, the complexity of the item format, the complexity of the target construct, and the number of items ultimately needed to be written. As Clark & Watson (2019) note, "No existing data-analytic technique can remedy item-pool deficiencies" (p. 4). Some practical tips for item writing include giving item writers plenty of time to make item writing quotas. Item writing can be mentally taxing and item writers can benefit from incubation, feedback, and a good night sleep!

Review of Item Features Research

When considering organizationally relevant constructs, measure developers must face some critical yet persistent questions related to item features. Some of these are whether to include reverse-coded items and subtle or ambiguous items and what type of response options should be chosen. In this section, I review some of the current research on these different areas.

Negative and reverse-coded items. This has been a persistent research question that has divided researchers for a long time. Advocates for negatively coded items suggest that they are important

in terms of dealing with response sets or careless responding and making sure that respondents are paying attention to items. If all items are keyed in the same direction, it may be difficult to untangle people who are high in the construct from people who are high in the acquiescence response set (i.e., people who tend to answer in the affirmative on all items regardless of attention). The logic is that keying some items in the opposite direction requires respondents pay extra attention to the content of items, which will force them to more carefully consider the content of the items (see DeVellis 2003, Weijters & Baumgartner 2012). However, there is plenty of research that shows that negatively coded items have significantly lower discrimination (Sliter & Zickar 2014) and that they often load on separate methods factors (in addition to a traditional construct factor) in confirmatory factor analysis (CFA) (Schriesheim & Eisenbach 1995). To complicate the matter, at least for some constructs, it has been proposed that the opposite end of a construct is itself its own construct. For affect, positive and negative affect have been thought to be independent, although correlated, constructs instead of just opposite ends of a bipolar continuum (Crawford & Henry 2004); similar results have shown optimism and pessimism to be independent yet correlated constructs (Chang et al. 1997).

The research on negatively worded items is simple yet complicated. If the desire is to maximize unidimensionality, items should all be scored in the same direction. Reverse-coded items tend to have smaller discrimination and introduce additional methods factors. These methods factors tend to explain small amounts of variance but having reverse-coded items will introduce nonconstruct variance. However, by including items all in the same direction, variance due to response sets may be masked as true score variance, thus artificially inflating reliability. Weijters et al. (2013) present an empirically tested model about reversed item bias that demonstrated three sources that can cause bias (acquiescence, careless responding, and confirmation bias) and they demonstrate that the negative effects of reverse-coded items can diminish when respondents pay careful attention to the content of items. Unfortunately, with much of the research being conducted now using low-motivation respondents via online recruitment methods such as Amazon's Mechanical Turk, our samples are likely becoming full of careless and low effort respondents. Another suggestion is to avoid using simple negations (e.g., including the modifier NOT in a stem) for reverse-coded items, as research has shown those items to be particularly confusing to respondents (see Swain et al. 2008).

Subtle or ambiguous items. Another feature that is debated within the scale development literature is whether to include subtle or ambiguous items. These are more traditional in the clinical literature than in organizational research, but the motivation for including subtle items is tempting. For example, the item "I am against giving money to beggars" (Gynther et al. 1979) is unclear in its intent. Respondents will be unlikely to understand the underlying construct being measured by subtle items and so their responses will be less influenced by social desirability or faking. The research on subtle or ambiguous items is unambiguous in one way: These items tend to perform worse psychometrically. Using item response theory (IRT), I with colleagues have found that these items have less discrimination (Zickar & Ury 2002, Min et al. 2018), whereas research on whether subtle items increase scale validity tends to not support that (see Hollrah et al. 1995). Subtle items are tempting, as respondents will provide more honest information. The same motivation is behind a recent push by some organizations (e.g., Google) to use brainteasers during interviews, such as "Estimate how many windows are in New York" (see Highhouse et al. 2019). These items are thought to throw off respondents; to get them to think spontaneously instead of relying on canned, planned answers to expected questions; and to provide deep insight into candidates. Unfortunately, those brainteasers tend to have little validity and introduce random noise into selection decisions.

This motivation to reduce faking or socially desirable responding is behind many attempted scale development innovations, such as social desirability scales, appropriateness indexes, response latencies, and bogus pipelines (see Zickar & Gibby 2006 for a history of faking and detection attempts). The history of this research in terms of operationally being able to detect fakers is fairly discouraging, given most techniques that exist have extremely high false alarm rates (i.e., falsely identifying honest respondents as fakers). The best research out there indicates that warning applicants not to fake is the best way to prevent such faking (e.g., Dwight & Donovan 2003, Landers et al. 2011). In addition, I have always thought that a key variable that would impact whether applicants fake on a personality test is the likelihood that they think people they eventually work with will remember their test scores. Most applicants likely think that their personality test scores will be forgotten once the hiring decision has been made, or perhaps will never be disclosed to coworkers. To the extent that applicants think that their scores might have bearing on their job, they may feel more accountable to be honest. For example, if an extremely introverted candidate thinks that faking being an extravert will ensure that in some future date her supervisor will expect extraverted-like behavior, that would likely be discouraging. All of this is a digression, although relevant, from the general conclusion in the literature that subtle items are likely to lead to poor psychometric functioning with little likely gain in validity.

Related to subtle items are double-barreled items. In these items, there are two qualifying statements within one particular item stem, and the respondent must answer the conjunction of these two stems. For example, one item may state "Employers should be allowed to use urine drug tests but not hair follicle tests." With traditional scale development, such items are judged to be poor items. People who disagree with this item may think neither urine testing nor hair follicle testing are okay, or they may think that both are okay. Because of this, these items tend to have low item-total correlations with classical test theory (CTT)-based methods. With the advent of unfolding IRT models (as detailed in the section Item Response Theory), such items (under certain conditions) are thought to be important because they can potentially measure moderate levels of the latent trait. For some situations, these items can be useful. For example, Zhang et al. (2015) used double-barreled items such as "clarifies work requirements, but does not micromanage work" to measure the construct of paradoxical leadership, which relates to the ability to manage competing and complex behaviors. In general, the best advice is to avoid double-barreled items in most cases because they confuse respondents. More research is needed on this issue, and Cao et al. (2015) provide another resource that should be considered for those working in the personality domain.

Response options. The preceding rationale has largely focused on the item stems, although the response scale is important and requires careful consideration. Most psychological measures have response formats that relate to either frequency of a particular event or behavior mentioned in the stem, or the response option indicates the degree of agreement or disagreement with a particular statement. There are two sets of literature dealing with response options. One line of literature investigates how the nature of the response options changes the psychological meaning of particular items, whereas the other literature focuses on practical differences between option features. In the former, Schwarz (1999) looked at how frequency options influenced how respondents interpreted items. For example, in the use of assessment of anger, if a frequency range is relatively short (e.g., last week) compared to relatively long (e.g., last year), the interpretation of what is being asked by the item changes. In the former, respondents will interpret the item to be focused on items related to minor incidences of anger. In the latter, respondents will tend to interpret the items to be asking about major anger events (see also Winkielman et al. 1998).

For agreement formats, one of the considerations is whether to have a middle option or not. Typical middle option labels include "*Neutral*," "*Neither agree nor disagree*," or "*?*". Some advocates

for the middle option argue that it useful for individuals who are truly ambivalent on a particular attitude; leaving out the middle option would incorrectly force a respondent to choose an attitude direction that masks their ambivalence. Others argue that it is best to remove a middle option because people are rarely completely neutral about a particular option—and so it is better to force individuals to choose a particular side. In addition, some advocates against a middle option argue that the middle option attracts both individuals who are neutral as well as individuals who do not understand the item. Finally, there has been some recent work that suggests that if using an underlying measurement model, the ideal point model, then one should avoid middle response options (Dalal et al. 2014). This model conceptually posits that individuals agree or disagree with an item stem based on the theoretical distance that the individual has from the item. They argue that providing a neutral or middle option is incompatible with the underlying theoretical model and should be eliminated. Dalal et al. (2014) demonstrated that items without a middle option fit an ideal point model better than identical items, differing only in that they have a middle option. In the item and scale evaluation section, I review the ideal point model in more detail. Dalal et al.'s (2014) research, however, is a good reminder that an underlying response model should be used not just to guide item analysis but also item development.

Although most organizational-based scales tend to use a small number of response option formats, there has been research on innovative formats. Vergauwe et al. (2017) proposed asking respondents for leadership items on whether they had "much too little" (scored −4) of a particular skill to "the right amount" (0) to "much too much" (+4). They purported that this format applies to curvilinear relations between a latent trait and response propensity. As with most innovations, however, until more research has been conducted this response format should be treated as experimental.

Forced-choice (FC) item formats have been around since the 1940s, although with advances in psychometric theory they are making a comeback. With FC items, respondents are given two statements, often of roughly equal social desirability, and then they are asked to choose their preference for one of those two options. A classic example item is "Would you rather (A) mop up a bucket of spilled maple syrup, or (B) climb a very tall mountain?" Someone high in risk aversion would choose option A, whereas someone high in risk-seeking would choose option B. Other items pair item stems from two different constructs, such as, "I would describe myself as (A) likes foreign cinema, or (B) somebody who worries a lot." The first stem would relate to openness to experience, whereas the second stem relates to neuroticism.

Although FC items are thought to reduce faking, they introduce other challenges, with ipsativity being the primary limitation. For the latter item, someone may choose A over B because she is low in openness to experience but even lower in neuroticism. Somebody else may choose A over B because he is high in openness to experience but even more extremely high in neuroticism. With traditional scoring techniques, we are stuck with ipsative data, meaning that if someone chooses A over B, we cannot say that they are higher in openness than somebody who chooses B (that would be a normative comparison). We can only say, for that individual, we suspect that their level of openness would be higher than their neuroticism. This ipsative type of data might be useful for developmental purposes, such as deciding which of the Big Five traits a particular employee could use coaching on. For most purposes, however, organizational researchers prefer normative data that allows individuals to make comparisons between each other. Traditional scoring of FC did not allow such comparisons to be made.

McCloy et al. (2005) proposed a creative use of IRT that allowed normative data to be extracted from FC items, assuming that a systematic pattern of item presentation was followed. Additional researchers have expanded on this methodology and have caused somewhat of a revival of FC items (see Brown & Maydeu-Olivares 2013, Stark et al. 2012).

Innovative Item Features

Even as technology has advanced, measure development often seems wedded to limitations imposed by previous technologies. For example, most tests still have a fixed number of response options that are presented in text format even as technology has provided us with a wider variety of options. Video-based items and options have been used for at least the past 20 years. Olson-Buchanan et al. (1998) developed a situation judgement of conflict resolution in which conflict-based scenarios were shown to respondents and they were then given a series of text-based options to choose from. Their research showed that the conflict-skills test was unrelated to cognitive ability. The thought was that by removing some of the reading component from traditional SJTs, assessments could have the overlap with cognitive ability minimized, a desirable result for applicant testing where adverse impact concerns persist. Malamut et al. (2011) describe Marriott International's development of a web-based assessment used to hire housecleaning staff, a group that often has low literacy skills. The Heart-of-House assessment uses pictures of sample hotel rooms and asks respondents to identify cleaning mistakes and errors, thus allowing for an assessment of traits such as disposition toward neatness that does not require much English reading ability. In addition, for items that are primarily text-based, they allow respondents to respond to text items or voice-over recordings of the text option. These assessments translate the information from a traditional text-based item into a more audiovisual experience, although they do not formally change the nature of what an item is. In addition, the response options remain primarily text-based.

In terms of using technology to transform response options, one approach has been to remove Likert scale (e.g., Strongly Disagree to Strongly Agree) options and replace them with a sliding scale that allows the respondent to choose their precise location and not be limited by discrete values. Couper et al. (2006) conducted an experiment comparing a graphical slider version with a radio-button scale to a traditional text-based option. Item means did not differ across formats, which was encouraging, although they found that missing data and response times were increased for the sliding scale. Their results suggest that there would be little to be gained with the additional graphical flexibility. In addition, their results suggest that we have shaped our respondents over years of practice on how to respond to particular item formats. Often, more complex options present confusion and difficulty to respondents, although it may also be that respondents can adapt as technology and item writing adapts.

The possibility of natural language algorithm scoring will surely entice developers of new scales to consider more open-ended options. With natural language scoring, individual words in a person's response are analyzed and compared to a preset dictionary and scored based on complex algorithms that can take in various complex contexts. There has already been a series of studies that link language usage to personality traits. Park et al. (2015) analyzed the Facebook posts of 66,732 users with natural language processing and found reasonable convergences between scores based on natural-language scoring and self-report assessments of the Big Five personality traits (*r*'s ranged from 0.38 to 0.46). Test developers may consider allowing for open-ended response options for particular items and scoring them with natural language processing software. For example, respondents may be asked to answer a series of questions such as "Tell me the most difficult decision you had to make in the last year" along with "What has been your proudest accomplishment?" Respondents may be asked to type in their responses or to read their answers into a device that records them and then transcribes the information into text to be analyzed. Answers could then be assessed based on correspondence to a dictionary keyed to score the language on specific traits or attitudes. This open-endedness may appeal to researchers who are concerned that our formats shape particular answers as well as others who may feel that it is easier to fake an answer

if it requires just circling a particular number or pushing a particular button, and that it may be harder to concoct a false narrative on the fly using specific language. A difficulty in this approach, however, is that the language requirements for obtaining enough information for a reliable assessment may preclude or limit its use.

ITEM AND SCALE EVALUATION

Once items have been written, the statistical fun part begins. At this point, different psychometric frameworks need to be considered and various item analysis techniques need to be considered. Ideally, these decisions will be made based on theoretical appropriateness, although as usual in the real world, pragmatic considerations should be acknowledged as well. In this section, I review in brief detail some of the main tenets of some of the primary statistical frameworks guiding item and scale evaluation. Given the nature of this review, I touch on the main points of each approach and then refer readers to more extensive sources. I then review how these techniques can be used to advance measure development.

Classical Test Theory

Most common statistical approaches to item and scale evaluation can be tied to CTT.

The basic tenet of CTT is quite simple:

$$X = T + E.$$

In this equation, X refers to an observed score, which generally in most analyses would refer to the scale score. T refers to the true score, and E refers to an error term. Both T and E are theoretical concepts that are often misunderstood, partly as a function of their labels. The simple description of T is that it is the expected value of an observed score. That is, this is the best guess for how a test taker will perform. I think of the brainwashing experiment as a way to conceptualize the true score. Pretend that all of a test taker's memories of an exam are erased (sort of like the movie Ground Hog Day) when he or she repeats the exam over and over again. The true score would be the mean of all of the test scores. Many people believe that T is a latent trait score that is directly related to the true value of the underlying trait and that it exists independently of the particular test. This is called the platonic interpretation of the true score and is incorrect. Ideally, we hope that the true score is related to some underlying latent trait of the construct that we are hoping to measure. But there are tests that have no meaningful underlying trait (e.g., add an arithmetic item, an extraversion item, and an item measuring your credit score), and there will still be an expected score for this scale.

The error term E is also commonly misunderstood, partially by the label of it. The simple definition of this term is that it is the difference between the true score and the observed score, $E = T - X$. Basically, E should be thought of as a residual term. Because both T and E are unobservable, certain assumptions need to be made to parse observed score variance into true score variance and error variance. This is where reliability theory comes in.

The choice of a reliability coefficient determines the meaning of the error term. If test-retest reliability is chosen, the error term will be related to differences in scores across two time periods. For example, a test taker may have had a headache at time 1 and did better at time 2 because the headache was gone. Or he or she may have guessed correctly at time 1, and at time 2 was less lucky. With internal consistency, the error term will be related to domain sampling error. For example, a test taker may have scored higher on a reading comprehension test because it asked about a

town he or she grew up near. He or she got that item correct (and hence had a higher score than expected) because of insider knowledge, not because of superior reading comprehension skills. The item responses will appear to be somewhat inconsistent as judged just by reading comprehension. With test-retest, the error term ignores all sources of variance due to domain sampling error, and with internal consistency, all time-related error is ignored. In the latter, a test taker may have a headache at time 1 and be unable to concentrate and thus item scores may all reflect a lower value of the trait, but because data are collected at only one time point, it is not apparent that this source of variance is actually error. The ignored sources of error get lumped into the true score term.

With CTT-based item analysis, some of the simplest analyses can be the most important, especially at early stages. Computing simple item-level and option-level descriptive statistics can help identify particular problems at early stages. For example, with knowledge tests an option frequency analysis, often called a distractor analysis, can indicate whether particular options are chosen frequently and which are rarely chosen. Options that are chosen rarely might be rewritten or even eliminated. Options that are not scored as the correct answer but that are chosen frequently might be evaluated to see if they are too close to the right answer. With Likert scale items, item means can provide a good sense of the items that might be extremely high or low in social desirability. The other classic CTT-based item statistic is the item-total (corrected) correlation, which is the CTT-based index of item discrimination. This statistic examines the relationship of the item-level score with the total score of all other items sans the item under consideration. In initial consideration, these statistics can be helpful in weeding out bad items and identifying items that are improperly scored or need to be reverse-coded. There are various guidelines given for appropriate levels of item-total correlations. From my experience, item-total correlations are most useful at an early stage in identifying items that are clearly missing the mark in terms of the overall construct. Eliminating items that have much smaller (or nonsignificant) item-total correlations than the rest of the items is the first step. Remember, too, however, that restriction of range can influence correlations, and so items that have extreme means (high or low) may have low item-total correlations even though those items may be important for other reasons.

There are various stages of item analysis. CTT-based techniques are often most useful in the beginning stages of measure development and can be useful with relatively low sample sizes given that the statistics being used, means and correlations, can be relatively well estimated with small sample sizes (e.g., approximately 75 or more). These techniques can be useful in refining the item pool so that the data are relatively clean and stable to be analyzed with more sophisticated techniques, such as exploratory factor analysis (EFA), CFA, and IRT, all techniques that require much larger sample sizes.

Exploratory factor analysis. The goal of EFA is to take a series of items and identify a specific number of factors that can best represent the underlying characteristics shared by items. This has been an extremely popular scale development tool throughout the history of psychometrics, though it has also been fraught with controversy, especially as more advanced methods such as CFA have been developed. The underlying mechanics of EFA are extremely complicated and require being well-versed in matrix algebra and so are beyond the scope of this review, but the basic foundation of EFA is the correlation matrix of items. In terms of the item level, within factor analysis, the observed score variance in a particular item can be broken down into communality and uniqueness. Communality refers to variance that is shared across items and may be due to a general factor that is shared across all items or more specific factors that are shared across a subset of items. The uniqueness refers to variance that is not shared with any other items within a scale. The end result of a factor analysis is a decision on the number of factors to retain as well as loadings that reveal which underlying factors are measured by individual items. Involved in a EFA is a series

of decisions that need to be considered, and I rely on Fabrigar et al.'s (1999) excellent review article to structure this section. They include study design, determining whether EFA is appropriate, type of estimation, selection of number of factors, and factor rotation as the decision points. I would add factor interpretation as an additional consideration point key for organizational researchers.

In terms of study design, one of the key considerations is the small size requirements. There are various heuristics that have been proffered such as Gorsuch's (1983) recommendation of at least 5 participants per item with a minimum of 100 total participants. Nunnally (1978) recommended 10 participants per item in the EFA. The preferred way of resolving these kinds of issues is the Monte Carlo simulation in which the true structure is known and then data are generated to conform to those structures and then analyzed to see under what conditions the right or known answer is obtained. MacCallum et al. (1999) found that the needed sample size to accurately recover a particular factor structure depends on the complexity of the solution as well and the number of items that measure each factor. In a case where there is high communality (i.e., 0.70 or higher, that is 70% of the item-level variance due to common factors) and there are at least three items per common factor, a sample size of 100 may be sufficient. In more complex cases, larger sample sizes will be needed. One challenge with these heuristics is that it is often difficult to determine in advance the complexity of the solution and the degree of communality within items. Unless the scale has been well-analyzed and refined, it is likely that sample sizes larger than 200 will be needed to have replicable results.

The second decision point of Fabrigar et al. (1999) is whether EFA is appropriate or another similar approach such as principal components analysis (PCA) should be used. PCA is similar to factor analysis in practice (in SPSS, it is easy to choose between the two as they are in the same factor command), although conceptually there are significant differences. In short, EFA reduces items to a set of underlying factors that are designed to best "explain" the underlying variance across items. PCA uses a set of principal components to try to understand all of the variance across items. Fabrigar et al. (1999) state it well: "The objective of PCA is to determine the linear combinations of the measured variables that retain as much information from the original measured variables as possible" (p. 275). There has been much hand-wringing over which is preferred. Practically speaking, both approaches tend to give similar answers (although there are times when the results do not converge). That is, items tend to group similarly on factors in EFA, and there are usually similar components in PCA analysis of the same data. From a scale development point of view, I much prefer EFA, although, because of the distinction of common and unique variance, which is not present in PCA. Related to whether and when to use EFA is the question of whether it should be used or whether CFA should be. After I review the CFA approach to measure development, I address that issue.

With respect to number of factors to extract, there are several considerations. Most reviews of EFA focus on statistical considerations, but I also think conceptual clarity needs to be considered as well. The scree plot is the most important piece of data used in identifying number of factors. Eigenvalues (a mathematical term that relates to the potency of each individual factor) are key for the scree plot. Mathematically, the value of the eigenvalue of a preceding factor must be larger or of the same size as the subsequent factor. That is the eigenvalue of factor 2 cannot be larger than the eigenvalue of factor 1. The scree plot marks the eigenvalues of each factor, making the drop in eigenvalues apparent as the factor numbers increase. The process of identifying the number of factors has progressed over the years. In the early days, the criterion was choosing factors with eigenvalues greater than 1.0, and in other cases, researchers were told to look for the bend in the factors—that is, find out there is a precipitous drop in the value of the eigenvalues and that after the bend there is minimal change from factor to factor. In the former, an arbitrary amount was chosen to identify factors, whereas in the latter, there are often multiple bends in a scree plot and

so it can be somewhat unclear. Parallel analysis adds some rigor to the factor identification process. In this case, the scree plot based on the data is compared to a scree plot based on comparative data (i.e., same sample size and same number of items) based on randomly generated data. Factors that have eigenvalues that are higher than the randomly generated data should be retained and the others should be ignored (see Hayton et al. 2004). Besides these statistical criteria, with EFA I find that conceptual coherence is a meaningful criterion, even if it is more subjective. Factors that result in factor loadings that make little conceptual sense are a good indication that this factor might be spurious and difficult to replicate.

In terms of types of rotation, there are several points to consider. In factor analysis, there are an infinite number of factor solutions (i.e., loadings, and factor correlations) that can fit the data. Researchers have developed variation rotation methods to impose order on the factor solutions and to provide the most meaningful factor loadings. Consider factor loadings to be regression coefficients linking the underlying trait to the response on a particular item. Most factor rotation techniques work to impose simple structure on the factor loading matrix, seeking solutions that result in either relatively high (in absolute value) loadings or loadings close to zero. Some rotation techniques seem to maximize a general factor and others are more likely to identify secondary factors. The main decision related to factor rotation is whether one wants orthogonal or oblique factors. Under orthogonal rotation, underlying factors are uncorrelated with each other, whereas under oblique rotation the factors are allowed to be correlated with each other, and the factor correlations are estimated. In general, there seems to be reasonable consensus that oblique rotations (e.g., promax, direct oblimin) are preferred for several reasons. The general argument is that hardly anything is truly uncorrelated in organizational research and so orthogonal rotations (e.g., varimax) force unnatural solutions upon data. In addition, the argument goes that if constructs are truly uncorrelated, oblique rotations will allow one to observe that, because in that case correlations between factors will be estimated to be zero. Despite these arguments, orthogonal rotations continue to be used because they often provide simple and conceptually clear factors (see Nunnally 1978). In general, however, I believe that oblique solutions should be used in that parsimony should not take precedence over accuracy.

Finally, I include factor interpretation as a key point in EFA. If done truly in an exploratory manner, EFA should be a tool of discovery. There are a lot of decision points that go into a factor analysis, and these decision points will impact the outcome of the analysis. The end goals of most EFAs, at least in the context of measure development, must be kept in mind. In most cases, an initial EFA is attempting to identify if the preconceived notions about construct are consistent with the data. For example, is a two-factor scale actually bi-dimensional? It may be that there are additional, unanticipated factors, or perhaps respondents did not distinguish between factors and there is only one meaningful factor in the data. Specific questions should be pursued with CFA, but in an initial analysis, an EFA might be appropriate to consider alternative solutions that could not even have been anticipated beforehand. One cannot test something in CFA that might not have even been considered. Also, key in EFA is factor interpretation. If you obtain a statistically meaningful factor that just cannot be interpreted, the question is what do you do with that factor? Typically, these factors are the factor retained last, and so one strategy is to rerun the EFA extracting one less factor and seeing if the results are more conceptually clear and clean. There are no right answers on what to do in this situation, but if EFA is used in a truly exploratory manner, these points should entice serious reflection and consideration. Sometimes, this reflection can result in insights about scale and its underlying construct(s). That is truly the goal of EFA.

Confirmatory factor analysis. CFA can be considered the next wave in factor analysis in that what happens in CFA is that researchers have an a priori structure that they believe underlies their data,

and CFA is able to test explicitly where the CFA fits the data or not. In general, EFA relies more on subjective judgement for final interpretation, whereas CFA relies on explicit statistical criteria for final interpretation. As I make clear in this section, however, there are still many subjective decisions that go into CFA.

The key to a CFA analysis is setting up an a priori model, which is really just drawing a series of arrows leading to and from circles and squares. With CFA, you identify a priori underlying latent traits (circles) and link observed items (squares) together via arrows. With most measures, the arrow goes from the latent trait to the item, documenting that the latent trait causes the item response. (There are some measures, called formative constructs, in which the areas go from the indicators to the latent trait; see Diamantopoulos & Siguaw 2006). In addition, there are residuals that relate to variance unexplained by any latent trait. Finally, there are double-headed arrows that specify that there is a correlation between items or constructs, although the direction of that relationship is unspecified. In scale development work, double-headed arrows are most often used to indicate that latent constructs are correlated with each other.

In terms of steps within a CFA, the first is model specification. This is the key step in measure evaluation and in which the most insight can be obtained. It is important that a lot of thought goes into this stage. One should identify the number of factors within the scale and identify which items load onto which factors. There are also many specific statistical considerations that need to inform a model. For example, each latent trait should have at least three indicators, although there may be two if there are correlated latent traits. If a model is improperly specified, the analysis may not converge or may give improper results. Most CFA software can give you advice on how to proceed in these situations. Perhaps the most important thing that can be done in CFA is the comparison of competing models. For example, you can test how well a one-factor model fits compared to a more complicated two-factor solution.

The next stage in the analysis is model estimation. There are various techniques available, although in most cases, researchers use maximum likelihood estimation (MLE). Cases in which one might consider other approaches to MLE include having extremely non-normal data and having many data with few response categories. In these cases, weighted least squares estimation may be considered.

The next stage is model evaluation. There is a whole literature on how to best evaluate CFA models (e.g., Jackson et al. 2009, Nye & Drasgow 2011). Common statistics that are suggested include the Root Mean Square Error of Approximation, Goodness of Fit Index, the Adjusted Goodness of Fit Index, Tucker-Lewis Index, the Comparative Fit Index, the Chi-square/df ratio, and the Standardized Root Mean Square Residual. Some of these indexes correct for complexity; that is, more complex models (i.e., more parameters to estimate), everything else equal, will fit data better because they have more flexibility. With increased flexibility, however, comes increased risk of capitalization on chance. Although there are many heuristics out there to suggest when there is good fit or not, one must remember what the goal of the scale development and evaluation goals are for which the analysis is being conducted. As mentioned earlier, CFA works best when testing alternative factor structures. In addition, CFA provides some important tools that can be useful in investigating issues related to measurement invariance.

One challenge in scale development for CFA is modifying the model to improve fit after models have been evaluated. In scale development contexts, there can often be significant increases in model fit by making small changes to the model, such as correlating a series of item residuals with each other, or allowing one item to cross-load on another factor in addition to the one originally hypothesized. There are differences in opinion on whether such changes should be allowed with CFA. Some people, the hardliners, argue that any model modifications that were based on data from the same sample in which the model was estimated should not be allowed. By modifying

based on data in the same sample, technically your approach is no longer confirmatory, but now retains an element of exploration. It is common for items that load on different factors to have a correlated residual. For example, two items that load on different personality traits might need a correlated residual term because they both contain the word exceptional within them. This shared variance between items might be due to some respondents having peculiar interpretations of the word exceptional and because both items include that word in the stem, those items may be more highly correlated than would otherwise be expected. If the goal is to maximize model fit, the correlated residual should be estimated and then reported (perhaps in a footnote in a manuscript). Perhaps in a further revision, the word exceptional is removed from one of the item stems and replaced with a synonym. The CFA approach can be a very important tool in the scale development process in that it helps researchers think clearly about the process of scale structure, and the fine-detailed metrics provided by CFA allow some level of analysis simply not possible with EFA- and CTT-based methods. A dogmatic approach to the confirmatory nature of CFA, in the context of scale development, is counterproductive in my opinion. As argued throughout, the process of scale development and evaluation is ongoing. What is learned in one data collection can be used to modify the model in a subsequent data collection.

Item response theory. The last of the remaining evaluation frameworks is IRT (e.g., Embretson & Reise 2000, Zickar 2012). At its core, IRT relates an underlying trait, commonly denoted by the Greek letter theta (θ), to the probability of choosing a particular answer or option, using a specified statistical model, usually referred to as the item response function (IRF). The IRT model chosen has parameters that need to be estimated. These parameter estimates form the heart of IRT scale evaluation. For dichotomous data, the most commonly used model has been the 3PL model, which is represented by the following equation:

$$P(u_i = 1|\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-1.7\,a_i(\theta - b_i)}}. \qquad 1.$$

As noted before, θ refers to the level of the latent trait for an individual, typically assumed to be normally distributed. In the 3PL model, each item is represented by three parameters (**Figure 1**).

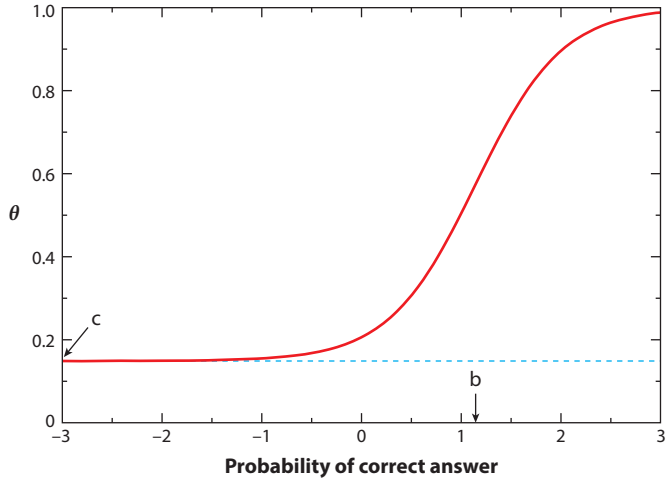

Figure 1

Item response function for 3PL model. Item characteristic curve (a = 1.351, b = 1.146, c = 0.149).

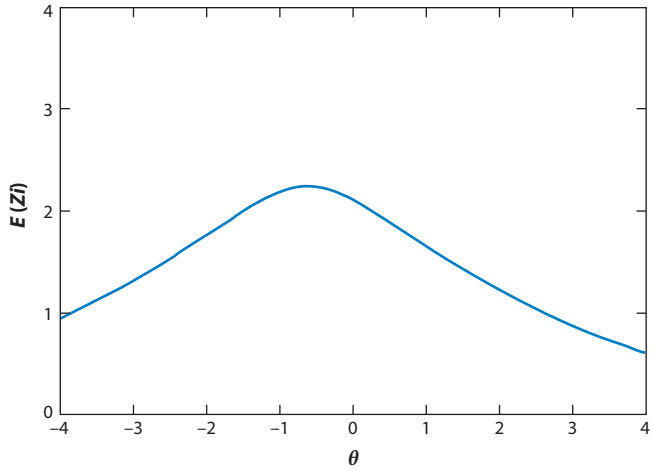

Figure 2

Item response function for generalized graded unfolding model. (Item 16: $\delta = 0.620$; $\alpha = 0.419$; $\tau = -3.420$, –2.141, 0.600, 3.087).

The *a* parameter represented the amount of discrimination that an item has, whereas the *b* parameter represents an item's location, and the *c* parameter represents the item's guessing propensity. When parameters are estimated, you can plot the IRF and see how θ relates to the probability of getting an item correct (or answering in the affirmative).

Models vary in terms of their flexibility and complexity. Although the 3PL is designed to model dichotomous items, there are many models designed to model polytomous (i.e., more than two options) response options. Among polytomous models, some assume ordinal- or interval-level scaling, whereas other models can handle nominal data. Most models assume unidimensionality, although other models allow for multiple dimensions of θ to model the item response process. Finally, most models assume a dominance approach, that is, that more of a θ always leads to an increase in the probability of answering an item correctly. Some models, however, assume an ideal point approach in that each item has an ideal point and the farther a θ is away from that ideal point, the probability of answering the item correctly diminishes, regardless of whether you have too much or too little θ. **Figure 2** provides an unfolding item as measured by the Generalized Graded Unfolding Model. This item measures attitudes toward drug testing in the workplace and asks the respondent to agree or disagree with the statement "Employers should be allowed to use urine drug tests but not hair follicle tests." Respondents who have extremely negative items about drug testing in the workplace will disagree with the item because they believe no drug testing should be allowed. Those who are extremely positive toward workplace drug testing will also disagree, but because they believe all types of testing should be allowed.

There are similar steps in an IRT analysis compared to the CFA approach. Models need to be chosen and then estimated to fit a particular data set. Next, fit is considered, although in this case, IRT modelers often focus on whether a particular item fits the chosen IRT model, which is different from the CFA approach, which tends to focus on more global issues of fit. One interesting feature of IRT is the item and scale information function. Information in IRT is a value that corresponds to the amount of uncertainty that is reduced in the estimation of θ by administering a particular item. Each item has separate information functions and the value of information varies by θ. A very difficult item might provide a high amount of information at the high end of θ, being

useful in distinguishing geniuses from everybody else. But that same item might not be useful in distinguishing between those below average in the trait compared to those who are average (both sets would be expected to get the item wrong). A nice feature of the item information functions is that they can be summed to form a test or scale information function. This function shows how the test provides information through the various ranges of θ and can be extremely useful in scale evaluation and modification. One can find which area of θ the test provides the most information for and which areas are most lacking. The desired shape of a test information function is dictated by the purpose of the test. For example, a test that is designed to identify children who need remedial help in school should have the most information at the lower end of θ, whereas a test designed to identify students for a gifted program should have lots of information at the high end of θ. Now if the same test was required to do both functions, as some intelligences tests do, that test should have a high level of information throughout the range of θ.

The information function is key to computerized adaptive testing (CAT), which chooses items to administer based on the amount of information that corresponds to an individual's theta estimate, an estimate that is updated after every item response. With CAT, each individual receives a test best chosen to provide precise measurement for that individual (see Wainer et al. 2000). In addition, the information function for scale reduction can be useful as well. In that case, one can eliminate items that provide little information throughout the range of θ that you decide is most important (see Russell et al. 2004). In this way, items can be eliminated that contribute least to the overall scale functioning.

WHEN TO CHOOSE BETWEEN THE VARIOUS TEST THEORY APPROACHES

The four frameworks CTT, EFA, CFA, and IRT all have important places in the toolbox of the scale developer and can be considered in many ways complementary approaches instead of competitors. Although it is tempting to think of more recent frameworks such as IRT and CFA as superior to the earlier frameworks of CTT and EFA, the reality is that none of the frameworks has been made redundant or obsolete by subsequent advances. The best particular tool for a particular purpose depends on the sample size requirements, the degree of formal structures of hypotheses that one has about an instrument, assumptions about the nature of responding, and the purpose of the analysis. I reveal some key points to consider for all four of these areas.

Sample Size

With small samples under 150, there is little that you can do besides a simple CTT-based item analysis, identifying items that have low item-total correlations or little variance in option responses. This is likely okay in initial stages of data collection if the goal is just to identify items that do not function at a minimal standard. Research shows that if you are looking to identify items with low discrimination, the correlation between item (corrected) total correlations tends to be extremely high with IRT-based item discrimination parameters (Ellis & Mead 2002). The same is likely with CFA and EFA factor loadings. With IRT, it is possible to conduct analyses with sample sizes of approximately 150 to 250, although in those cases the model must be quite simplified. With small samples, one would be limited to the 1PL or perhaps the 2PL model. More complex models such as polytomous IRT models and multidimensional models would require sample sizes closer to approximately 500 cases. For EFA, as mentioned previously, sample size requirements depend on the complexity of the underlying model, as does CFA in which the number of parameters estimated depends precisely on the model. More complex models require larger data sets. Most

scale development efforts start small and then get progressively larger data sets as steps progress. This seems reasonable, given that in early stages, the most important thing is to identify items that simply do not work well, either due to poor writing, unanticipated confusion, or contamination.

Degree of formal hypothesized structure. If researchers have a well-defined hypothesized structure of the instrument that they are assessing, then measures with more confirmatory possibilities should be used. If you have developed a new three-factor measure of organizational politics, you should test that model using CFA or another approach that allows you to reject or quantify degree of model fit. Alternatively, one might use multidimensional IRT to test whether items load on the hypothesized structure. Even in these more confirmatory situations, however, I urge researchers to use the tool's powers to investigate alternative models and explore deviations from the expected model. As Spector et al. (2014) noted, the early days of organizational research were largely atheoretical, although the pendulum has swung so that with analytic approaches that can explicitly test theory, there has been a more thorough embrace of deductive methods. They argue, however, that inductive approaches still warrant attention. In the scale development context, I argue that a degree of openness to alternative solutions is important and that understanding when and why models fail to work can lead to interesting discoveries that may make a scale more important.

Assumptions about the nature of responding. Different test frameworks make different assumptions about the nature of response process used by the respondent. Some of the most exciting work in psychometrics has come in recent years in terms of opening up the possibilities in this domain. The work on ideal point responding provides a significant contrast to the traditional view of item response process that more of a trait is better in terms of responding to an item correctly or in an affirmative manner. This notion has existed ever since Thurstone (1927) but it took until the mid-1990s for an IRT-based approach to operationalize the model in a way that was convenient for researchers to test (see Drasgow et al. 2010). Nearly every other psychometric approach assumes that more is better. For example, CTT assumes that the relation between T and X is linear as does EFA and CFA (although there are some nonlinear factor analyses approaches), and most IRT models assume a more-is-better approach as well. As technology examples, there are a diversity of models that have been proposed, including many IRT models that allow for different classes of individuals to use different response models, thus wedding latent class approaches with IRT (see von Davier & Carstensen 2007). This model has been used to uncover different types of personality test faking within a particular sample (Zickar et al. 2004) and identifying individuals using different response sets on attitudinal items (Hernández et al. 2004). Researchers should think deeply on what is the best underlying model for a particular construct, then they should choose an appropriate model, and finally they should test to see whether that particular model is appropriate.

Purpose of the analysis. The purpose of the analysis is perhaps the most important of these four factors, and the purpose of the analysis should of course guide all aspects including the collection of data and the choice of measurement model. As noted earlier, if the goal is just to pick items that will result in a high reliability, CTT is fine. However, if you have specific goals in maximizing measurement precision in a particular range of the underlying trait, then IRT would be preferable. If the goal is to develop a test that functions similarly across different groups (e.g., works similarly across men and women), then an approach that allows an explicit test of that would be preferable. IRT or CFA would be the preferred way to test the measurement equivalence of the scale. If the goal is to determine how one scale relates to other scales (e.g., is the construct measured by your new scale significantly different than that measured by an existing test) then a CFA approach

Table 2 Measure evaluation frameworks

Framework	Summary
CTT	Advantages: simplest of all test theories, statistical techniques include item total correlations and internal consistency reliability coefficients, works well with small sample sizes, is useful in early stages of scale development
	Disadvantages: no way to determine whether the model fits the data, assumes linear relation between true score and observed score
EFA	Advantages: useful in inductive research to identify constructs being measured by a test, which can be helpful in early stages of research to help refine the construct definition
	Disadvantages: often produces unreplicable results, especially with small sample sizes; many subjective decisions that may impact outcomes
CFA	Advantages: provides specific evaluations of fit that can be used to test the relevance of competing factor models, useful once the basic structure of a test has been clarified
	Disadvantages: results often hypersensitive to item wording factors
IRT	Advantages: provides detailed statistics on the functioning of individual items, allows understanding of the response process, can use many sophisticated psychometric tools such as computerized adaptive testing
	Disadvantages: requires large sample sizes and complex statistical programs

Abbreviations: CFA, confirmatory factor analysis; CTT, classical test theory; EFA, exploratory factor analysis; IRT, item response theory.

would be most preferable. Finally, if your goal is to develop an adaptive test that allows different items to be administered to different test takers, the flexibility and precision of IRT are necessary. **Table 2** summarizes some of the key advantages and disadvantages of each of the 4 psychometric frameworks.

CONCLUSIONS

Many of the approaches to measure development approach the process as a statistical problem with a set of heuristics and guidelines that need to be followed strictly. Throughout this review, I argue that statistical techniques are an important part of scale development, but equally important are the initial conceptual development work and item design efforts. In addition, scale development at its best can engage a spirit of discovery that will result not just in measures that have good statistical properties but also allow for rich theoretical understandings.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Brown A, Maydeu-Olivares A. 2013. How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychol. Methods* 18(1):36–52

Buckles S, Walstad WB. 2008. The national assessment of educational progress in economics: test framework, content specifications, and results. *J. Econ. Educ.* 39(1):100–6

Cao M, Drasgow F, Cho S. 2015. Developing ideal intermediate personality items for the ideal point model. *Organ. Res. Methods* 18(2):252–75

Chang EC, Maydeu-Olivares A, D'Zurilla TJ. 1997. Optimism and pessimism as partially independent constructs: relationship to positive and negative affectivity and psychological well-being. *Personal. Individ. Differ.* 23(3):433–40

Clark LA, Watson D. 2019. Constructing validity: new developments in creating objective measuring instruments. *Psychol. Assess.* In press

Couper MP, Tourangeau R, Conrad FG, Singer E. 2006. Evaluating the effectiveness of visual analog scales: a web experiment. *Soc. Sci. Comput. Rev.* 24(2):227–45

Crawford JR, Henry JD. 2004. The Positive and Negative Affect Schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample. *Br. J. Clin. Psychol.* 43(3):245–65

Dalal DK, Carter NT, Lake CJ. 2014. Middle response scale options are inappropriate for ideal point scales. *J. Bus. Psychol.* 29(3):463–78

DeVellis RF. 2003. *Scale Development: Theory and Applications*, Thousand Oaks, CA: SAGE. 2nd ed.

Diamantopoulos A, Siguaw JA. 2006. Formative versus reflective indicators in organizational measure development: a comparison and empirical illustration. *Br. J. Manag.* 17(4):263–82

Drasgow F, Chernyshenko OS, Stark S. 2010. 75 years after Likert: Thurstone was right!. *Ind. Organ. Psychol.* 3(4):465–76

Dwight SA, Donovan JJ. 2003. Do warnings not to fake reduce faking? *Hum. Perform.* 16(1):1–23

Ellis BB, Mead AD. 2002. Item analysis: theory and practice using classical and modern test theory. In *Blackwell Handbooks of Research Methods in Psychology: Handbook of Research Methods in Industrial and Organizational Psychology*, ed. SG Rogelberg, pp. 324–43. Malden, MA: Blackwell Publ.

Embretson SE, Reise SP. 2000. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum

Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. 1999. Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* 4(3):272–99

Fives H, DiDonato-Barnes N. 2013. Classroom test construction: the power of a table of specifications. *Pract. Assess. Res. Eval.* 18(3):1–7

Gierl MJ, Haladyna TM, eds. 2013. *Automatic Item Generation: Theory and Practice*. New York: Routledge

Gorsuch RL. 1983. *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Gynther MD, Burkhart BR, Hovanitz C. 1979. Do face-valid items have more predictive validity than subtle items? The case of the MMPI Pd scale. *J. Consult. Clin. Psychol.* 47(2):295–300

Haladyna TM, Rodriguez MC. 2013. *Developing and Validating Test Items*. New York: Routledge

Hayton JC, Allen DG, Scarpello V. 2004. Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. *Organ. Res. Methods* 7(2):191–205

Hernández A, Drasgow F, González-Romá V. 2004. Investigating the functioning of a middle category by means of a mixed-measurement model. *J. Appl. Psychol.* 89(4):687–99

Highhouse S, Nye CD, Zhang DC. 2019. Dark motives and elective use of brainteaser interview questions. *Appl. Psychol.* 68(2):311–40

Hollrah JL, Schlottmann RS, Scott AB, Brunetti DG. 1995. Validity of the MMPI subtle items. *J. Personal. Assess.* 65(2):278–99

Jackson DL, Gillaspy JA Jr., Purc-Stephenson R. 2009. Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychol. Methods* 14(1):6–23

Landers RN, Sackett PR, Tuzinski KA. 2011. Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *J. Appl. Psychol.* 96(1):202–10

MacCallum RC, Widaman KF, Zhang S, Hong S. 1999. Sample size in factor analysis. *Psychol. Methods* 4(1):84–99

Malamut A, Van Rooy DL, Davis VA. 2011. Bridging the digital divide across a global business: development of a technology-enabled selection system for low-literacy applicants. In *Technology-Enhanced Assessment of Talent*, ed. NT Tippins, S Adler, pp. 267–92. San Francisco, CA: Jossey Bass

McCloy RA, Heggestad ED, Reeve CL. 2005. A silk purse from the sow's ear: retrieving normative information from multidimensional forced-choice items. *Organ. Res. Methods* 8(2):222–48

Min H, Zickar M, Yankov G. 2018. Understanding item parameters in personality scales: an explanatory item response modeling approach. *Personal. Individ. Differ.* 128:1–6

Nunnally JC. 1978. *Psychometric Theory*. Hillsdale, NJ: McGraw-Hill. 2nd ed.

Nye CD, Drasgow F. 2011. Assessing goodness of fit: Simple rules of thumb simply do not work. *Organ. Res. Methods* 14(3):548–70

Olson-Buchanan JB, Drasgow F, Moberg PJ, Mead AD, Keenan PA, Donovan MA. 1998. Interactive video assessment of conflict resolution skills. *Pers. Psychol.* 51(1):1–24

Park G, Schwartz HA, Eichstaedt JC, Kern ML, Kosinski M, et al. 2015. Automatic personality assessment through social media language. *J. Personal. Soc. Psychol.* 108(6):934–52

Russell SS, Spitzmüller C, Lin LF, Stanton JM, Smith PC, Ironson GH. 2004. Shorter can also be better: the abridged job in general scale. *Educ. Psychol. Meas.* 64(5):878–93

Schriesheim CA, Eisenbach RJ. 1995. An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *J. Manag.* 21(6):1177–93

Schwarz N. 1999. Self-reports: how the questions shape the answers. *Am. Psychol.* 54(2):93–105

Shaffer JA, DeGeest D, Li A. 2016. Tackling the problem of construct proliferation: a guide to assessing the discriminant validity of conceptually related constructs. *Organ. Res. Methods* 19(1):80–110

Sliter KA, Zickar MJ. 2014. An IRT examination of the psychometric functioning of negatively worded personality items. *Educ. Psychol. Meas.* 74(2):214–26

Spector PE, Rogelberg SG, Ryan AM, Schmitt N, Zedeck S. 2014. Moving the pendulum back to the middle: reflections on and introduction to the inductive research special issue of Journal of Business and Psychology. *J. Bus. Psychol.* 29(4):499–502

Stark S, Chernyshenko OS, Drasgow F, White LA. 2012. Adaptive testing with multidimensional pairwise preference items: improving the efficiency of personality and other noncognitive assessments. *Organ. Res. Methods* 15(3):463–87

Swain SD, Weathers D, Niedrich RW. 2008. Assessing three sources of misresponse to reversed Likert items. *J. Mark. Res.* 45(1):116–31

Thurstone LL. 1927. A law of comparative judgment. *Psychol. Rev.* 34(4):273–86

Vergauwe J, Wille B, Hofmans J, Kaiser RB, Fruyt FD. 2017. The too little/too much scale: a new rating format for detecting curvilinear effects. *Organ. Res. Methods* 20(3):518–44

von Davier M, Carstensen CH. 2007. *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*. New York: Springer

Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ. 2000. *Computerized Adaptive Testing: A Primer*. New York: Routledge

Weijters B, Baumgartner H. 2012. Misresponse to reversed and negated items in surveys: a review. *J. Mark. Res.* 49(5):737–47

Weijters B, Baumgartner H, Schillewaert N. 2013. Reversed item bias: an integrative model. *Psychol. Methods* 18:320–34

Winkielman P, Knäuper B, Schwarz N. 1998. Looking back at anger: Reference periods change the interpretation of emotion frequency questions. *J. Personal. Soc. Psychol.* 75(3):719–28

Zhang Y, Waldman DA, Han YL, Li XB. 2015. Paradoxical leader behaviors in people management: antecedents and consequences. *Acad. Manag. J.* 58(2):538–66

Zickar MJ. 2012. A review of recent advances in item response theory. In *Research in Personnel and Human Resources Management*, ed. JJ Martocchio, A Joshi, H Liao, pp. 145–76. Bingley, UK: Emerald Group Publ.

Zickar MJ, Gibby RE. 2006. A history of faking and socially desirable responding on personality tests. In *A Closer Examination of Applicant Faking Behavior*, ed. RE Griffith, MH Peterson, pp. 21–42. Charlotte, NC: Information Age Publ.

Zickar MJ, Gibby RE, Robie C. 2004. Uncovering faking samples in applicant, incumbent, and experimental data sets: an application of mixed-model item response theory. *Organ. Res. Methods* 7(2):168–90

Zickar MJ, Ury KL. 2002. Developing an interpretation of item parameters for personality items: content correlates of parameter estimates. *Educ. Psychol. Meas.* 62(1):19–31