A Typology of Threats to Construct Validity in Item Generation

Lucy R. Ford
Saint Joseph's University

Terri A. Scandura
University of Miami

We review the literature on item generation and develop a typology of common threats to construct validity related to poor item generation practices. The typology is based on the literature on item generation, which suggests that these are practices to avoid in the construction of survey items. The threats are categorized as scale-centered or context-centered, and might occur in either item construction or in the meaning drawn by the respondent. By wording items to avoid these pitfalls researchers may be able to improve the construct validity of their studies.

INTRODUCTION

There has been much attention paid in the research methods literature to the issues of construct validity of measures; however, treatment of the issue of item generation is scant, and contained to basic coverage in survey research methods textbooks (Dillman, 1978; Dillman, Smyth, & Christian, 2014; Fowler, 1993). Moreover, there is little research on the content validity of survey items, or the extent to which individual items measure the content domain of the construct (Schriesheim, Powers, Scandura, Gardiner, & Lankau, 1993). It should also be noted that it is generally accepted that content validity is a necessary precondition for construct validity (Schriesheim et al., 1993). However, it seems obvious that poorly written items will result in poor psychometric properties of measures. Thus, we feel that it is time to look upstream in our construct validation process to the genesis of the survey items themselves.

The costs of poor measurement in organizational research are substantial. Survey items that make no sense to survey respondents may result in high levels of non-response and/or response bias. Participants may even be alienated by emotion-laden questions that are difficult to answer (Narins, 1999). DeVellis (2003) notes that "…researchers often "throw together" or "dredge up" items and assume they constitute a suitable scale" (p. 11). Such practices result in poor measures and results that may not be replicable. Poor measures result in a lack of construct validity and the inability of statistical procedures to detect statistical significance. And yet, even if significance is detected, the lack of good measurement may result in inappropriate conclusions being made. For example, while authors often reference the exchange of benefits between a leader and a follower, a commonly used measure (LMX-7) does not directly measure exchange per se (Bernerth, Armenakis, Feild, Giles, & Walker, 2007). We also need to be concerned about whether the items in each scale mean the same thing to each respondent (Fowler, 1993). Another concern is that there is a proliferation of scales in the literature, purportedly measuring different things,

but appearing to differ little in content (Shaffer, DeGeest, & Li, 2016). An example of this is the surprising finding that two measures which should be quite different, engagement and job burnout, overlap considerably. Meta-analytic findings demonstrated that dimension-level correlations between burnout and engagement are high, and burnout and engagement dimensions exhibit a similar pattern of association with correlates (Cole, Walter, Bedeian, & O'Boyle, 2012). Moreover, this study found that controlling for burnout in meta-regression equations substantively reduced the effect sizes associated with engagement. Clearly the content of items is fundamental to the advancement of organizational research, yet little attention has been paid to the practice of item generation, either in the literature or in practice. As MacKenzie, Podsakoff and Podsakoff (2011) note "it is important to remember that the effectiveness of any content adequacy assessment technique is only as good as the definitions of the construct (and the items) that are developed by the researcher in the first place" (p.306).

The purpose of this paper is to review the literature on item generation and develop a typology of common threats to construct validity due to poor item generation practices. This typology should prove useful to the researcher who wishes to deliberately avoid these threats to construct validity in their scale development practice. While other authors (i.e. Dillman et al., 2014; Hinkin, 1995; DeVellis, 2003; Hardy & Ford, 2014) have addressed item generation in their work, none of the sources consulted for development of this typology have provided authors with a reasonably comprehensive list of threats to construct validity, with clear prescriptive descriptions allowing avoidance of unnecessary error.

Item Generation: A Brief Overview

Content validity is a necessary but not sufficient condition for a measure to have construct validity. And of course, item generation is an essential step in the construct validity process (Hinkin, 1995). Content validity has been defined as the degree to which a particular measure reflects a specific intended content domain, while construct validity is concerned with "the extent to which a particular measure relates to other measures consistent with theoretically derived hypotheses concerning the concepts…that are being measured" (Carmines & Zeller, 1991). In many cases the issue of construct validity is handled quite succinctly: "Use published measures." Thus, we have many dissertations and research studies that employ only measures that have been used in prior research and have purportedly acceptable psychometric properties. While this advice appears sound, there are some problems with this advice that are apparent upon closer scrutiny. First, many published and commonly used measures have threats to construct validity due to poorly written items. Many measures in use may therefore not measure the construct they are purported to measure, or they may be indistinguishable empirically from scales intended to measure similar but distinct constructs, thereby casting doubt on the construct validity of the scale. Second, measures may be well written, but time bound, such that "...as years go by, the contents of some test items may become dated. Therefore, periodic revisions may be required to keep test item wording current" (Warner, 2008, p. 871).

Third, if only published measures are employed, then the field does not move forward in new and relevant directions. This is analogous to "reshuffling" the same old deck of cards with the same tired constructs (do we really need another study of job satisfaction and organizational commitment no matter how good the measures are purported to be?) Our field is in need of new ideas, new constructs and (of course) new measures that represent them well, in order to extend the nomological net beyond currently existing knowledge. In addition, revision of existing scales that have less than desirable construct validity is necessary. For example, the concept of empowerment emerged as a new organizational phenomenon in the 1980s. Spreitzer (1995), while borrowing from and adapting existing measures, carefully developed and validated a measure of empowerment that could be used in organizational research, using a deductive approach. She specified four dimensions of empowerment and presented a preliminary nomological network of relationships between empowerment and organizational concepts. This work moved the field forward so that new studies might be conducted on an emerging organizationally-relevant topic. New measures are thus often necessary to advance organizational knowledge. For example, Brown, Trevino and Harrison (2005) developed a new measure of ethical leadership. None existed as it was a relatively new construct. In their paper, they describe a rigorous process conducting seven studies, each with unique

samples, in order to develop, evaluate and refine the scale, and then in the last three studies, to establish discriminant and predictive validity of the measure. Such thoughtful scale development practices exist in the literature, but tend to be the exception rather than the rule.

Where do survey items come from?

As noted by Schwab (1980) content validity is essential to the item generation process. There are generally two ways to develop survey items: Deductive and inductive. Deductive scale development begins with a classification scheme or typology prior to the writing of the items and subsequent data collection. This approach requires a thorough understanding of the construct of interest and a complete review of the literature so that consonant theoretical definitions are generated prior to the development of items. A clear operational definition grounded in theory is required for the deductive approach. In some cases, subject matter experts are asked to review the items to ensure that the items accurately reflect the domain of the construct. The second method is the inductive method. The inductive approach requires little theory prior to the writing of the items. The research generates measures from individual items. With this approach, researchers usually develop items by asking a pretest sample of respondents to describe their work experiences, reactions to something that is happening in the workplace, or aspects of their own behavior at work. For example, the researcher might ask, "Tell me about your reactions to your organization undergoing downsizing." Responses might include "The downsizing is causing me stress," or "I have seen others lose their jobs, and I am worried that I will lose mine." Content analysis is then used to classify responses to develop items on the basis of these responses.

To examine practices employed to establish content validity, Hinkin (1995) reviewed a sample of 75 studies published between 1989 and 1993 that developed new measures. With respect to item generation procedures, he found that most studies (83%) employed deductive methods, 11% were inductive and 6% employed a combination of methods. Deductive approaches using a classification scheme or typology are by far the most common, and this may be due to the emphasis on strong theory in the major journals. Inductive approaches were far less common. They are used when there is little or no theory to guide item writing.

We are not advocating one method over the other. Deductive procedures may be more appropriate when there is well-established theory to guide the writing of items. Inductive processes might be necessary when there is a new line of inquiry for which the input of respondents is needed. Ideally, both processes should be used in a reflexive manner, however only a small percentage of studies in Hinkin's review employed both. While it is not the purpose of our study to review the approaches employed to generate items, we are in agreement with DeVellis (2003) that the process is often less systematic than is desirable.

Item Generation Threats to Construct Validity

In the process of generating items it is important to keep in mind the importance of construct validity. There are a number of potential threats to construct validity that relate to the wording of the items themselves, which can be avoided in the item generation process. We present a typology of these threats and define each based upon the research methods literature. We based this typology on the literature on item generation, which suggests that these are practices to avoid in the construction of survey items.

Ambiguity

Despite admonitions regarding ambiguity that relate back to a classic paper on the construction of attitude scales, the problem persists: "Above all, regardless of the simplicity or complexity of vocabulary or the naiveté or sophistication of the group, each statement *must avoid every kind of ambiguity."* (Likert, 1973, p.91). We defined item ambiguity as: *Any statement that is confusing, vague, or otherwise subject to multiple interpretations.* Respondents cannot answer such items validly because they may not understand what the item is requiring them to respond to (Oppenheim, 1992). For example, an item that asks respondents about their intent to quit their job in the near future might be interpreted by some as next week, and by others as being in the next six months (Hardy & Ford, 2014). It is more effective, therefore,

to provide temporal anchors for time-bound words. Another example might be the use of the word "organization". An item that asks respondents whether their "organization treats them well" may invoke different referents in the minds of individual respondents, if the term "organization" is not clearly defined. For example, for one employee, their organization might be represented by their individual manager, while for another it might be represented by the nameless, faceless executives at a distant headquarters. Clearly the interpretations of the word are very different for these two respondents, with the researcher having no way of knowing what their referent was when responding. In addition, clearly the term "organization" will mean something different to an executive at the corporate office than it will to a manual laborer at a remote company site. Alreck and Settle (1985) suggest that the only sure way to avoid such ambiguity is to ensure that every questionable word or phrase in every item is thoroughly checked to ensure that they invoke the same meaning in every respondent.

Leading Questions

We considered leading questions to be *items which lead the respondent to believe there is a correct answer* (Narins, 1999). For example, "Smoking causes cancer. How much do you smoke?" In this case, the additional information given in the question might prompt the respondent to underreport the amount they smoke. In addition, we considered items to be leading if they contain implicit assumptions, where the researcher assumes that some component must be true (Saris & Gallhofer, 2007). For example, an item such as "When did you last talk with your mentor about your career plan?" assumes that the respondent has a mentor, and/or that they have talked with that person about their career plan. The respondent may answer it even if they don't have a mentor, or they haven't had that conversation, and thus the incidence of the behavior would be over-reported. Experiments conducted by Swann and his colleagues found that when a questioner asks a leading question of a respondent, observers use their knowledge of conversational rules to infer that the questioner had an evidentiary basis for the question (Swann, Giuliano, & Wegner, 1982). Respondents treated leading questions as conjectural evidence as implied by the question. This research also found that when respondents answer leading questions, they are in fact misled by the answers because they want to cooperate with the person posing the question, and may misrepresent their own personality traits. Leading questions typically arise as the result of the use of non-neutral language as demonstrated by the examples given above (Penwarden, 2015).

Double-Barreled

This is the threat to construct validity most frequently mentioned in texts and articles on item generation as a problem to avoid (e.g. Converse & Presser, 1986; Hinkin, 1998; Oppenheim, 1992; Saris & Gallhofer, 2007). We define double-barreled questions as *those items that ask multiple questions within a single statement* (e.g. Are you happy and well-paid?) In this case the respondent may be happy or well-paid but they would be responding to only part of the question (Hinkin, 1998). Saris defines this issue as a situation where "two simultaneously opposing opinions are possible" (2007, p. 87). Stated differently, when a respondent endorses a double-barreled item, it cannot be known which part is the source of the positive (or negative) response to the question. Such items may also be confusing to survey respondents since they are not sure exactly what the question is.

Reverse Coding

In some instances, *statements that reflect the polar opposite of the construct* (e.g. for job satisfaction: I am unhappy in my job) are employed in survey measures. While using both positive and negative items in a scale is intended to avoid an acquiescence bias (DeVellis, 2003), research on negatively-worded items has indicated that these may produce response bias (DeVellis, 2003; McGee, Ferguson, & Seers, 1989; Merritt, 2012). This phenomenon has been studied empirically (Cordery & Sevastos, 1993) and theoretically (Marsh, 1996). In many cases it appears that the reverse coded items end up loading on a different factor than the one intended, and researchers are advised to pay careful attention to dimensionality when using reverse scored items. For example, Magazine, Williams and Williams (1996) demonstrated the method effects resulting from reverse coded items in their examination of a measure of

Affective and Continuance Commitment. If reverse coded items are used explicitly only to test for careless responding, there are other methods that might be used that would not result in the error associated with a reverse coded item. For example, the researcher might use items that all respondents might be expected to answer in the same manner if they are paying attention, such as "I was born on February 30th" (Huang, Curran, Kenney, Poposki, & DeShon, 2011).

Negative Wording

Some authors (e.g. DeVellis, 2003) use this term to refer to reverse coded items. Others, however, use the term to mean *statements that include negative words* (such as "not") irrespective of whether they are measuring the construct or the polar opposite of the construct. It is this latter meaning to which we are referring, as the former is encoded in the previous term, reverse coding. The negative meaning is conveyed by the choice of words, such as "restrict", "not", and "control". Negatively worded statements may bias responses since they may result in lack of endorsement (Converse & Presser, 1986) due to such concerns as social desirability bias or the potential that certain words might trigger a negative emotional reaction in participants (Design, 2017). Belson (1981) discusses items with a negative element as a factor which might impair a respondent's understanding of the item. Another concern raised by Harrison and colleagues (1993) is that a block of negatively worded items can have a cognitive carryover effect, resulting in biased responses to subsequent neutral questions.

Double Negatives

Statements in which a double negative occurs may be particularly confusing for respondents (e.g., "Pay for performance is not an unjust policy"). Double-negatives exist when *two negative words are linked in the same item* (Converse & Presser, 1986; Oppenheim, 1992). Here the intent of the researchers might be to have a positive evaluation of pay for performance but the respondent will perhaps indicate a negative response due to the wording (e.g. (Hinkin, Tracey, & Enz, 1997). Another example is when a double negative is more subtle and it ends up in a survey item. Consider this Agree/Disagree item: Please tell me whether you agree or disagree with the following statement about supervisors in organizations: Supervisors should not be required to supervise direct reports when they are leaving work in the parking lot. A person may agree that supervisors should not be required to supervise direct reports in the parking lot. But the Disagree side gets confused because it means "I do not think that supervisors should *not* be required to supervise direct reports when they are in the parking lot. This might happen when researchers don't read aloud and listen to all the questions in a series (Converse & Presser, 1986). Anyone who has seen double negatives in an exam question will recognize the problems with this particular item wording concern. For example, we may ask students to select the correct answer to the question: "Which of the following is the least unlikely result of poor leadership?"

Jargon

This refers to statements in which technical jargon appears that might not be understood by all respondents (Oppenheim, 1992). Jargon is the use of *organizational textbook terms or buzzwords from the industry*. For example, an item that asks respondents about emotional intelligence or social learning might not be understood by all respondents. In the information technology field, a question about cloud computing may confuse as there is no universal definition of the term, and different organizations may have a different view of its meaning. And while the terms stereotype and prejudice have very distinct meanings for scholars, the average survey respondent is unlikely to differentiate between the two. This particular threat to construct validity is relatively easy to avoid, though, by using the plainest and most commonly used words that accurately convey the intended meaning.

Colloquialisms

This refers to statements in which *slang appears – phrases that may be misunderstood by non-native English speakers or respondents who have not kept up with the latest in slang* may be a threat to content validity. In addition, items may be interpreted differently by respondents from different areas of the US,

different countries, or respondents from different industries (e.g., a statement that asks if your supervisor would "go to bat" for you). Oppenheim (1992) warns of the possibility of alternative meanings for colloquialisms. For example, "passing the buck" may mean to hand off responsibility in the US, but might be interpreted to mean handing over money in other parts of the world (Dunham & Smith, 1979)! Hardy and Ford (2014) highlighted the different interpretations of words and phrases by those from different geographic regions, such as the word momentarily which means "for a moment" to a British English speaker, and "in a moment" to an American English speaker. This latter example is not strictly a colloquialism, except to the degree that it creates the same threat to construct validity as is caused by the use of colloquialisms.

Acronyms

Abbreviations of words may represent a concern as not all respondents may be familiar with the abbreviations (Oppenheim, 1992). For example, acronyms like LMX (Leader-Member Exchange) or TQM (Total Quality Management) may or may not have any meaning or the same meaning for all respondents to a survey. Further, while organization-specific acronyms may be known to most organizational members, the use of such acronyms without definition might pose a challenge to new organizational members who have not yet learned all the internal language. Furthermore, there are plenty of acronyms that have multiple meanings. For example, CD may refer to a Computer Disk, or a Certificate of Deposit.

Prestige Bias

Prestige bias is defined as *statements that would prompt the respondent to agree with high-status experts*. Statements that begin with "All experts agree that…" might produce a response of agreement rather than disagreement even if the respondent disagrees with the "experts", as they believe they "should" align their response with those who "know best". Another example would be "four out of five CEOs surveyed endorse strategic planning. Do you?" Prestige bias has been shown to have pervasive effects in respondents. For example, prestige bias has been shown to affect the peer review process in academic journals (Lee, Sugimoto, Zhang, & Cronin, 2013). Since this effect is present among scholars who should be aware of such bias, we can expect that it will also influence the average survey respondent. Most authors who write about prestige bias conflate it with social desirability bias. We believe, however, that they are distinct forms of bias, as it may be possible for a question to include prestige bias, and skew responses in the opposite direction to that which would be socially desirable.

Social Desirability Bias

This possible threat is similar to prestige bias, but without the invocation of a prestigious other. Such statements may prompt *bias towards a socially acceptable response*. For example, "Would you save a drowning person?" In this case, people are reluctant to answer a question honestly because they are trying to appear socially or politically correct. The motivation to "save face" or impress others might alter the manner in which a respondent answers a question. Due to the potential for this bias, many researchers use a scale in their surveys to detect socially desirable responding (DeVellis, 2003). Moorman & Podsakoff (1992) reported results of a meta-analytic review of 33 studies that examined the relationships between social desirability response sets and organizational behavior constructs. Their findings showed that social desirability, as traditionally measured in the literature through social desirability measures (e.g.Crowne, 1960; Paulhus, 1989), is significantly (although moderately) correlated with several widely used constructs in organizational behavior research. Social desirability bias can create serious construct validity problems, in addition to being cited as one of the causes of self-report bias in organizational research when all of the data is collected from the same source (Donaldson & Grant-Vallone, 2002).

Acquiescence Bias

In the literature there are two distinct definitions of acquiescence bias. The first is when the scale, taken as a whole, encourages the respondent to respond positively to all of the items (e.g. DeVellis, 2003).

The second refers to a statement that is *likely to prompt a "yes" response in respondents*. For example, asking respondents if they read the newspaper every day may produce a "yes" response. However, in reality, they may only have time to read the paper a few times a week or on Sundays. Alreck and Settle (1985) suggest that the question should not give any indication of which is the "preferred" response, in order to avoid this bias. In the newspaper example, acquiescence bias could be avoided by asking the respondent "How often do you read a newspaper?"

Summary

The preceding sections reviewed a number of potential threats to construct validity that emanate from the writing of survey items themselves. There has been recent attention in the literature on organizational research on the examination of the items themselves. For example, Carpenter et al. (Carpenter, Son, Harris, Alexander, & Horner, 2016) conducted a meta-analytic study at the item level of analysis on a commonly-used scale of task performance and organizational citizenship behavior, and found that several of the items did not perform in the initial validation of the measures. This study underscores the need for researchers to be more diligent in item generation and the construction of measures at the outset to ensure that measures perform as claimed by researchers in the original publication. This is consistent with the recent measure-centric approach to understanding method variance advocated by Spector et al. (2017) in which more attention should be paid to developing a theory of the measure. To add to this discussion, we have organized the threats to construct validity due to item generation into the following typology.

A TYPOLOGY OF THREATS TO CONSTRUCT VALIDITY

After developing the list of threats to construct validity, the authors discussed the list and it became clear that the threats can be categorized into four broad categories, which can be represented in a 2x2 matrix as shown in Table 1. Scale-centered threats (as opposed to context- centered threats) are those items which are inherently flawed in and of themselves. For example, most texts that address construct validity in item development recommend that items not be double-barreled. This reflects the content of the item itself. On the other hand, context-centered threats are those where the item may be inherently well-written, and not contain any scale-centered threats to construct validity, but because of the context, the item is misinterpreted by the participant. For example, there is no problem with using acronyms in an item if every participant is clear exactly what that acronym means in the context in which it is being used. On the other hand, if an acronym has alternative meanings known to the participants, then it is possible they will respond with the wrong one in mind.

TABLE 1
MATRIX OF THREATS TO CONSTRUCT VALIDITY

	Scale-centered	Context-centered
Item construction	Reverse coding Negative wording Double negatives	Colloquialisms Acronyms Jargon
Item meaning	Double-barreled Aquiescence	Ambiguity Leading questions Prestige bias Social desirability bias

On the other axis of the matrix are item construction and item meaning. Item construction refers specifically to the structure of the item itself. For example, it has been shown numerous times that negatively worded items create a bias (DeVellis, 2003; McGee et al., 1989) that affects the validity of results. This is a very specific problem within an item and is easily avoided. Conversely, item meaning refers to the potential interpretation on the part of the participant. For example, a double-barreled item

such as "I am happy with my pay and benefits" may elicit responses from participants in which they are focusing specifically on their pay, or on their benefits, or on both. This, of course, renders comparison among participants problematic.

DISCUSSION

In the future, we need to conduct more rigorous construct validity studies whenever we develop new scales or adapt existing ones. It is quite common for a researcher to "throw together" some items for a construct he/she needs to measure (DeVellis, 2003, p. 11), include them in a survey, and establish their construct validity through the use of exploratory/confirmatory factor analysis prior to testing the hypotheses in the study. DeVellis (2003, p. 86) suggests that all items should be reviewed by experts prior to use. He notes that the experts serve three purposes; to rate the extent to which they believe each item measures the construct of interest, what you have potentially failed to include, and most relevant to this paper, the clarity and conciseness of the items. Hardy and Ford (2014) go a step further in recommending that a sample of participants be asked what items mean to them in order to understand how respondents are receiving the items. Relying solely on academic experts may well result in more flawed items than if we ask respondents what the items mean to them. Ren and colleagues found that under some circumstances those with greater expertise may have stronger cognitive biases than those with less expertise (Ren, Simmons, & Zardkoohi, 2017). This concern is shared by Hardy and Ford (2014) who found that research methodologists made more mistakes in interpreting survey instructions than did regular study participants. We believe that engaging in construct validation with this typology as a guide will result in researchers uncovering the subtler problems in generated items.

Many existing measures could potentially be improved by rewording items to avoid the pitfalls outlined in this paper, and evaluating the context in which the survey is being administered, in order to assess the likelihood that context centered threats to construct validity are a concern.

Experts with methodological training are likely to "catch" many item problems, particularly those that are scale centered, as they are related to the construction of the item itself, without concern for context. It has been our experience, based on our work in this area, that context centered threats to construct validity are more difficult for researchers to identify. In particular they often do not see an item as ambiguous, as their reaction to it is based on their own particular interpretation of the word(s), based on their own life experience. It is not always easy for the researcher to step out of their own "mental box" and imagine what other potential interpretations could be made of an item. A potential remedy would be to have multiple experts write out a description of what the item means to them, and in the event of an item referring to the organization, a group, etc., what their referent is. These descriptions could then be compared for consistency. Hardy and Ford (2014) demonstrated that respondents to surveys often interpret items differently, and that asking a sample of participants to describe the meaning of each item is a useful mechanism for uncovering miscomprehension.

If it is clear that respondents are not all interpreting the referent in the same manner, then it would behoove researchers to clearly delineate the appropriate referent in the instructions, in order to attempt to ensure that at least most of the respondents are thinking about it in the intended manner. This not a perfect solution as not all survey respondents read the instructions (Hardy & Ford, 2014). As organizational researchers seem to rarely, if ever, provide information in their published papers about the instructions that were provided to survey participants, it is difficult to assess whether any given study has been affected by this particular threat to construct validity. Perhaps it is time that we started reporting instructions in our methods sections, in an effort to ensure that measures are comparable across studies.

It can be useful to include participants from different geographical regions, different organizations, and with different native languages, in order to uncover other context-centered threats to construct validity in the items. For example, asking participants from different geographic regions to explain what they understand items to mean should uncover colloquialisms, as some participants will be less likely to understand them correctly.

This paper reviewed the threats to construct validity based upon item generation. It seems that many of the problems encountered with common method variance, and other problems emanating from poor measurement might be alleviated if more time were taken to construct the items in measurement scales from the outset. While we don't advocate specifically for the use of deductive or inductive methods to gain insight into item content, we do note that deductive methods are far more common in the literature, and perhaps a combination of the two approaches would be more helpful. By alerting researchers to the threats to construct validity in our typology, we hope that researchers will take more trouble to follow our recommendations and produce improved measures for business research, by using our recommendations in conjunction with existing guidelines on such things as scale length. These guidelines for developing items, used in conjunction with existing work on concerns about scale length and statistical methods for empirical evaluation of the generated scales, should result in improved scales in our literature.

ACKNOWLEDGEMENT

This work was supported by the Michael J. Morris Grant for Scholarly Research, Saint Joseph's University, and Board on Faculty Research and Development Summer Research Grant, Saint Joseph's University.

REFERENCES

Alreck, P. L., & Settle, R. B. (1985). *The survey research handbook*. Homewood, IL: Richard D. Irwin, Inc.

Belson, W. A. (1981). *The design and understanding of two survey questions*. Aldershot, Hants, England: Gower Publishing Co. Ltd.

Bernerth, J., Armenakis, A., Feild, H., Giles, W., & Walker, H. (2007). Leader-Member Social Exchange (LMSX): Development and validation of a scale. *Journal of Organizational Behavior,* 10.1002/job.443, 979-1003. doi:10.1002/job.443

Brown, M. E., Treviño, L. K., & Harrison, D. A. (2005). Ethical leadership: A social learning perspective for construct development and testing. *Organizational Behavior and Human Decision Processes, 97*(2), 117-134. doi:10.1016/j.obhdp.2005.03.002

Carmines, E. G., & Zeller, R. A. (1991). *Reliability and vaidity assessment*. Newbury Park: Sage Publications.

Carpenter, N. C., Son, J., Harris, T. B., Alexander, A. L., & Horner, M. T. (2016). Don't forget the items: Item-level meta-analytic and substantive validity techniques for reexamining scale validation. *Organizational Research Methods, 19*(4), 616-650. doi:10.1177/1094428116639132

Cole, M., Walter, F., Bedeian, A., & O'Boyle, E. (2012). Job burnout and employee engagement: A meta-analytic examination of construct proliferation. *Journal of Management, 38*(5), 1550-1581. doi:10.1177/0149206311415252

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire* (10.4135/9781412986045). Newbury Park, CA: Sage Publications, Inc.

Cordery, J. L., & Sevastos, P. P. (1993). Responses to the original and revised job diagnostic survey: Is education a factor in responses to negatively worded items? *Journal of Applied Psychology, 78*(1), 141-143. doi:0021-9010.78.1.141

Crowne, D. P. M., D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349-354. doi:10.1037/h0047358

Design, Q. (2017). Questionnaire design.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage. Publications, Inc.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: John Wiley.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business & Psychology, 17*(2), 245-260. doi:10.1023/a:1019637632584

Dunham, R. B., & Smith, F. (1979). *Organizational Surveys: An internal assessment of organizational health*.

Fowler, F. J. (1993). *Survey Research Methods* (2nd ed.). Newbury Park, CA: Sage Publications, Inc.

Hardy, B., & Ford, L. R. (2014). It's not me, it's you: Miscomprehension in surveys. *Organizational Research Methods, 17*(2), 138-162. doi:10.1177/1094428113520185

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. *Journal of Applied Psychology, 78*(1), 129-140. doi:10.1037/0021-9010.78.1.129

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

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104-121. doi:doi: 10.1177/109442819800100106

Hinkin, T. R., Tracey, J. B., & Enz, C. A. (1997). Scale construction: Developing reliable and valid measurement instruments. Retrieved from http://scholarship.sha.cornell.edu/articles/613

Huang, J. L., Curran, P. G., Kenney, J., Poposki, E. M., & DeShon, R. P. (2011). Detecting and deterring insufficient effort responding to surveys. *Journal of Business & Psychology, 27*, 99-114.

Lee, C., Sugimoto, C., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the Association for Information Science and Technology, 64*(1), 2-17. doi:10.1002/asi.22784

Likert, R. (1973). The method of constructing an attitude scale. In S. Houston, J. Schmid, R. Lynch, & W. Duff (Eds.), *Methods and Techniques in Business Research* (pp. 90-95). New York: MSS Information Corporation, Ardent Media.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrationg new and existing techniques. *MIS Quarterly, 35*(2), 293-295.

Magazine, S., Williams, L., & Williams, M. (1996). A confirmatory factor analysis examination of reverse-coding effects in Meyer and Allen's affective and continuance commitment scales. *Educational and Psychological Measurement, 56*(2), 241-250.

Marsh, H. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifacts? *Journal of Personality and Social Psychology, 70*(4), 810-819. doi:10.1037/0022-3514.70.4.810

McGee, G. W., Ferguson, C. E., & Seers, A. (1989). Role conflict and role ambiguity: Do the scales measure these two constructs? *Journal of Applied Psychology, 74*, 815-818. doi:10.1037/0021-9010.74.5.815

Merritt, S. M. (2012). The two-factor solution to Allen and Meyer's (1990) affective commitment scale: Effects of negatively worded items. *Journal of Business and Psychology, 27*(4), 421-436. doi:10.1007/s10869-011-9252-3

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology, 65*(2), 131-149. doi:10.1111/j.2044-8325.1992.tb00490.x

Narins, P. (1999). *Write more effective survey questions*. Retrieved from
Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement* (New ed.). New York: St. Martin's Press.
Paulhus, D. L. (1989). *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding - Version 6. Unpublished manuscript.*
Penwarden, R. (2015). 5 common survey question mistakes that'll ruin your data. Retrieved from https://www.surveymonkey.com/blog/2015/02/11/5-common-survey-mistakes-ruin-your-data/
Ren, R., Simmons, A., & Zardkoohi, A. (2017). Testing the effects of experience on risky decision making. *American Journal of Management, 17*(6).
Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research* (10.1002/9780470165195). Hoboken, NJ: John Wiley & Sons, Inc.
Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving Construct Measurement In Management Research: Comments and A Quantitative Approach for Assessing the Theoretical Content Adequacy of Paper-and-Pencil Survey-Type Instruments. *Journal of Management, 19*(2), 385. doi:10.1177/014920639301900208
Schwab, D. P. (1980). Construct validity in organizational behavior. In B. M. Staw & L. L. Cummings (Eds.), *Research in Organizational Behavior* (Vol. 2). Greenwich, CT: JAI Press, Inc. Shaffer, J. A.,
DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation: A guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods, 19*(1), 80-110. doi:10.1177/1094428115598239
Spector, P. E., Rosen, C. C., Richardson, H. A., Williams, L. J., & Johnson, R. E. (2017). A new perspective on method variance: A measure-centric approach. *Journal of Management, 014920631668729*. doi:10.1177/0149206316687295
Spreitzer, G. M. (1995). Psychological empowerment in the workplace: Dimensions, measurement and validation. *Academy of Management Journal, 38*(5), 1442. doi:10.2307/256865
Swann, W., Giuliano, T., & Wegner, D. (1982). Where leading questions can lead: The power of conjecture in social interaction. *Journal of Personality and Social Psychology, 42*(6), 1025-1032. doi:10.1037//0022-3514.42.6.1025
Warner, R. M. (2008). *Applied Statistics: From bivariate to multivariate techniques*. Thousand Oaks, CA:Sage.