American Psychological Association

Journal of Applied Psychology

© 2019 American Psychological Association
0021-9010/19/$12.00

2019, Vol. 104, No. 10, 1243–1265
http://dx.doi.org/10.1037/apl0000406

Content Validation Guidelines: Evaluation Criteria for Definitional Correspondence and Definitional Distinctiveness

Jason A. Colquitt, Tyler B. Sabey, Jessica B. Rodell, and Edwyna T. Hill
University of Georgia

Several reviews have been critical of the degree to which scales in industrial/organizational psychology and organizational behavior adequately reflect the content of their construct. One potential reason for that circumstance is a tendency for scholars to focus less on content validation than on other validation methods (e.g., establishing reliability, performing convergent, discriminant, and criterion-related validation, and examining factor structure). We provide clear evaluation criteria for 2 commonly used content validation approaches: Anderson and Gerbing (1991) and Hinkin and Tracey (1999). To create those guidelines, we gathered all new scales introduced in *Journal of Applied Psychology*, *Academy of Management Journal*, *Personnel Psychology*, and *Organizational Behavior and Human Decision Processes* from 2010 to 2016. We then subjected those 112 scales to Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches using 6,240 participants from Amazon's Mechanical Turk with detailed, transparent, and replicable instructions. For both approaches, our results provide evaluation criteria for *definitional correspondence*—the degree to which a scale's items correspond to the construct's definition—and *definitional distinctiveness*—the degree to which a scale's items correspond more to the construct's definition than to the definitions of other orbiting constructs.

Keywords: content validity, measurement, construct validity, scale development

Supplemental materials: http://dx.doi.org/10.1037/apl0000406.supp

Reviews in industrial/organizational psychology and organizational behavior continue to point to cases where published scales do not adequately sample the content associated with a construct. For example, Stone-Romero (1994) described scales in the affect, strain, and job attitude realms where items either failed to sample relevant content or instead sampled irrelevant content. He concluded "As a result, tremendous amounts of time, effort, and other resources have been expended on research that has poor conceptual and methodological underpinnings" (p. 175). More recently, Aguinis and Vandenberg (2014) argued that inadequate attention to content validation was one of a number of ways that "organizational science researchers begin their data analysis journey with a losing hand" (p. 590). Similarly, Aguinis and Edwards (2014) observed "there is no shortage of studies in which the correspondence between constructs and measures is tenuous" (p. 149). Such criticisms point to problems with *content validation*—the methodological process of gauging the degree to which scale items adequately sample the universe of content associated with a construct (Cronbach, 1990; Nunnally, 1978).

One potential reason that the fields of industrial/organizational psychology and organizational behavior continue to struggle with such issues is that content validation methods are utilized less frequently than other validation methods. Discussions of scale development note that a number of methods can be used to support the validity of inferences made when using a scale (Hendrick, Fischer, Tobi, & Frewer, 2013; Hinkin, 1995, 1998; MacKenzie, Podsakoff, & Podsakoff, 2011). Aside from content validation, those include examining reliability, convergent, discriminant, and criterion-related validation, and testing factor structure. As shown in Table 1, content validation has been discussed in the pages of *Journal of Applied Psychology* much less frequently than some of those other validation methods. Indeed, it is difficult to imagine a new scale being introduced without some discussion of reliability and factor structure. It is easy to imagine an article failing to include discussion of content validation.

The purpose of our article was to provide clear evaluation criteria and detailed instructions for two of the more common content validation methods in industrial/organizational psychology and organizational behavior: Anderson and Gerbing (1991) and Hinkin and Tracey (1999). Both approaches yield results relevant to *definitional correspondence*—a term used here to reflect the degree to which a scale's items correspond to the construct's definition. Both approaches also yield results relevant to *definitional distinctiveness*—the degree to which a scale's items correspond more to the focal construct's definition than to the definitions of other orbiting constructs. Currently there are some ambiguities in how to execute these two approaches and how to interpret the results that flow from them. We believe that making their execution and interpretation more clear could facilitate a more frequent use of the tools. Such clarity would also allow

This article was published Online First April 4, 2019.

Jason A. Colquitt, Tyler B. Sabey, Jessica B. Rodell, and Edwyna T. Hill, Department of Management, University of Georgia.

Correspondence concerning this article should be addressed to Jason A. Colquitt, Department of Management, University of Georgia, C210 Benson Hall, 630 South Lumpkin Street, Athens, GA 30602. E-mail: colq@uga.edu

Table 1
Prevalence of Validation Keywords in Journal of Applied Psychology

Validation keyword	Number of hits
Construct validation	502
Content validation	60
Reliability	9,540
Convergent validation	34
Discriminant validation	203
Criterion-related validation	68
Factor structure	1,540

Note. Searches were conducted within Google Scholar's advanced search feature on November 9, 2018. The validation keyword was entered into the "with the exact phrase" field for anywhere in the article, with *Journal of Applied Psychology* entered into the "Return articles published in" field.

reviewers to more critically judge the quality of new scales introduced to the literature.

To do so, we gathered all new scales published in four industrial/organizational psychology and organizational behavior outlets—*Journal of Applied Psychology*, *Academy of Management Journal*, *Personnel Psychology*, and *Organizational Behavior and Human Decision Processes*—from 2010 to 2016. There were 56 articles in that time frame that introduced at least one new scale, with many articles introducing multiple scales. In total, the 56 articles introduced 112 new scales to the literature. Those 112 scales are listed alphabetically in Table 2 and labeled as "focal scales" (the table also includes two "orbiting scales," which will be explained in a subsequent section). We should note that our benchmarking effort was limited to Likert-style scales of explicit constructs—approximately interval response scales where participants rate attitudes, cognitions, beliefs, traits, and behaviors in a nonimplicit, nontacit manner. Those types of scales were the focus of Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches and should therefore be considered a key boundary condition for our work and our resulting guidelines. We should also note that neither Anderson and Gerbing's (1991) approach nor Hinkin and Tracey's (1999) approach focuses on deficiency—the degree to which items fail to capture some part of the content universe (Cronbach, 1990; Nunnally, 1978). Our evaluation criteria are therefore only relevant to a subset of the concerns associated with content validation, a point we return to in the Discussion section.

Two Content Validation Approaches

Anderson and Gerbing (1991)

Anderson and Gerbing (1991) introduced their content validation approach for a practical reason. They noted the inefficiency created when scholars collected scale data, only to see the items load on unintended constructs in a confirmatory factor analysis. They therefore proposed a pretesting methodology that could be used to foreshadow how well items would perform in such analyses. Judges would be given a set of multiple construct definitions and a set of items. They would then be asked to read the items and sort them into the appropriate construct definition. Two indices—the proportion of substantive agreement (p_{sa}) and the substantive-validity coefficient (c_{sv})—would then be calculated as shown below, where N is the total number of judges, n_c is the number who sorted the item correctly, and n_o is the maximum number of times an item was sorted into any other construct in the set. The p_{sa} statistic ranges from 0 to 1, achieving the latter value when all judges classify an item correctly. The c_{sv} statistic ranges from -1 to 1, achieving the former value when no judges classify an item correctly and all do so incorrectly and the latter value when all judges classify an item correctly and none do so incorrectly.

$$p_{sa} = n_c / N$$

$$c_{sv} = (n_c - n_o) / N$$

In describing their approach, Anderson and Gerbing (1991) made reference to an earlier approach introduced by Lawshe (1975) in a selection context. Lawshe (1975) focused on the degree to which overlap existed between performance on a test item and effective job performance. His approach asked subject matter experts to classify the content of each item as either essential, useful but not essential, or not necessary for effective job performance. A statistic similar to those in Anderson and Gerbing's (1991) approach was introduced that would attain a perfect value when all judges viewed an item as essential for job performance. Despite that similarity, Anderson and Gerbing's (1991) approach differed from Lawshe's (1975) in two critical respects. First, the more general notion of a construct definition replaced the more specific focus on relevance for job performance. Second—and key to the generation of our evaluation criteria—Anderson and Gerbing (1991) eschewed the use of subject matter experts in favor of naïve judges. They wrote ". . . rather than being experts, judges in the pretest samples should be representative of the main study sample and population of interest," arguing in favor of "the capability of respondents with no formal training in psychology to make judgments about the relationships of items drawn from existing scales to the particular traits they reflect" (p. 734).

Anderson and Gerbing (1991) referred to both of their statistics as indicators of "substantive validity"—which they viewed as an item-level analog of the more scale-level "content validity." In an effort to be more specific in our jargon, we will refer to p_{sa} and c_{sv} as indicators of definitional correspondence and definitional distinctiveness. Thus, higher values reveal that a scale's items correspond more closely to the construct's definition. Higher values also reveal that a scale's items correspond more closely to the construct's definition than to the definitions of other orbiting constructs. It is important to note that the use of orbiting constructs relative to a focal construct creates a higher bar than a simple comparison against random classifications. Such values reveal items that are characteristic of—and diagnostic of—their intended construct, and useful for distinguishing that construct from other orbiting constructs.

Anderson and Gerbing (1991) demonstrated their approach using five personality constructs—energetic, impulse purchaser, thrill-seeking, avoids complexity, and unreflective—each of which was measured with seven items. Two samples of judges were employed, with each sample consisting of 20 undergraduates. The judges were shown the definitions of the five constructs on a sheet of paper, presented in random order. For example, one definition was presented as "**Thrill Seeking.** The *thrill-seeking*

Table 2
Alphabetical Listing of Focal and Orbiting Scales

Focal Scale	Orbiting Scale 1	Orbiting Scale 2
Avoidant Leader Behaviors (Gelfand, Leslie, Keller, & de Dreu, 2012)	Supervisor Initiating Structure (Stogdill, 1963)	Subjective Stress (Motowidlo, Packard, & Manning, 1986)
Bottom-Line Mentality (Greenbaum, Mawritz, & Eissa, 2012)	Supervisor Initiating Structure (Stogdill, 1963)	Achievement-Oriented Leadership (Indvik, 1985)
Calling (Dobrow & Tosti-Kharas, 2011)	Job Satisfaction (Cammann et al., 1983)	Job-Related Affective Well-Being: Activated Positive Arousal (Van Katwyk, Fox, Spector, & Kelloway, 2000)
Career Satisfaction (Seibert, Kraimer, Holtom, & Pierotti, 2013)	Job Satisfaction (Cammann et al., 1983)	Workplace Status (Djurdjevic et al., 2017)
Citizenship Fatigue (Bolino, Hsiung, Harvey, & LePine, 2015)	Role Overload (Bacharach, Bamberger, & Conley, 1990)	Subjective Stress (Motowidlo et al., 1986)
Collaborative Leader Behaviors (Gelfand et al., 2012)	Participative Leadership (Arnold, Arad, Rhoades, & Drasgow, 2000)	Supervisor Support (Oldham & Cummings, 1996)
Communicating High Expectations by Supervisor (Wang & Howell, 2010)	Supervisor Initiating Structure (Stogdill, 1963)	Participative Leadership (Arnold et al., 2000)
Competence Certainty (Mayer, Thau, Workman, Van Dijke, & De Cremer, 2012)	Self-Esteem (Rosenberg, 1965)	Workplace Status (Djurdjevic et al., 2017)
Constructive Voice (Maynes & Podsakoff, 2014)	In-Role Behavior (Williams & Anderson, 1991)	Civic Virtue (Podsakoff, MacKenzie, Moorman, & Fetter, 1990)
Contingent Reactions to Supervisor (Colquitt, Long, Rodell, & Halvorsen-Ganepola, 2015)	Satisfaction with Supervisor (Spector, 1985)	Trust in Supervisor (Brockner, Siegel, Daly, Tyler, & Martin, 1997)
Coworker Career Advancement (Colbert, Bono, & Purvanova, 2016)	Coworker Support (Susskind, Kacmar, & Borchgrevink, 2003)	Coworker Satisfaction (Spector, 1985)
Coworker Emotional Support (Colbert et al., 2016)	Workplace Friendships (Nielsen, Jex, & Adams, 2000)	Coworker Satisfaction (Spector, 1985)
Coworker Friendship (Colbert et al., 2016)	Coworker Support (Susskind et al., 2003)	Coworker Satisfaction (Spector, 1985)
Coworker Giving to Others (Colbert et al., 2016)	Meaning (Spreitzer, 1995)	Coworker Satisfaction (Spector, 1985)
Coworker Knowledge Utilization (Sung & Choi, 2012)	Coworker Satisfaction (Spector, 1985)	Coworker Support (Susskind et al., 2003)
Coworker Personal Growth (Colbert et al., 2016)	Coworker Support (Susskind et al., 2003)	Coworker Satisfaction (Spector, 1985)
Coworker Project Qualities (Long, Baer, Colquitt, Outlaw, & Dhensa-Kahlon, 2015)	Workplace Friendships (Nielsen et al., 2000)	Coworker Satisfaction (Spector, 1985)
Coworker Star Qualities (Long et al., 2015)	Coworker Support (Susskind et al., 2003)	Coworker Satisfaction (Spector, 1985)
Coworker Task Assistance (Colbert et al., 2016)	Coworker Support (Susskind et al., 2003)	Coworker Satisfaction (Spector, 1985)
Customer Unethical Behavior (Greenbaum, Quade, Mawritz, Kim, & Crosby, 2014)	Machiavellianism (Jonason & Webster, 2010)	Subjective Stress (Motowidlo et al., 1986)
Decision Control/Autonomy Supervisor Behaviors (Zhang, Waldman, Han, & Li, 2015)	Participative Leadership (Arnold et al., 2000)	Supervisor Initiating Structure (Stogdill, 1963)
Defensive Voice (Maynes & Podsakoff, 2014)	Interpersonal Deviance (Bennett & Robinson, 2000)	Psychological Withdrawal (Lehman & Simpson, 1992)
Desire for Power (Belmi & Pfeffer, 2016)	Extraversion (Donnellan, Oswald, Baird, & Lucas, 2006)	Machiavellianism (Jonason & Webster, 2010)
Destructive Voice (Maynes & Podsakoff, 2014)	Interpersonal Deviance (Bennett & Robinson, 2000)	Psychological Withdrawal (Lehman & Simpson, 1992)
Distance/Closeness Supervisor Behaviors (Zhang et al., 2015)	Supervisor Consideration (Stogdill, 1963)	Supervisor Humility (Owens, Johnson, & Mitchell, 2013)
Distraction Stigmas (Rodell & Lynch, 2016)	Psychological Withdrawal (Lehman & Simpson, 1992)	Role Overload (Bacharach et al., 1990)
Distributive Fairness Monitoring (Long, Bendersky, & Morrill, 2011)	Feedback Seeking (Ashford, 1986)	Active Coping (Carver, Scheier, & Weintraub, 1989)
Distributive Justice Clarity (Qin, Ren, Zhang, & Johnson, 2015)	Promotion Satisfaction (Spector, 1985)	Pay Satisfaction (Spector, 1985)
Dominating Leader Behaviors (Gelfand et al., 2012)	Achievement-Oriented Leadership (Indvik, 1985)	Participative Leadership (Arnold et al., 2000)

(*table continues*)

Table 2 (*continued*)

Focal Scale	Orbiting Scale 1	Orbiting Scale 2
Duty to Codes Orientation (Hannah, Jennings, Bluhm, Peng, & Schaubroeck, 2014)	Conscientiousness (Donnellan et al., 2006)	Prosocial Identity (Grant, Dutton, & Rosso, 2008)
Duty to Members Orientation (Hannah et al., 2014)	Agreeableness (Donnellan et al., 2006)	Prosocial Identity (Grant et al., 2008)
Duty to Mission Orientation (Hannah et al., 2014)	Task Proactivity (Griffin, Neal, & Parker, 2007)	Individual Initiative (Moorman & Blakely, 1995)
Emphasizing Group Identity by Supervisor (Wang & Howell, 2010)	Supervisor Consideration (Stogdill, 1963)	Supervisor Vision Articulation (Conger & Kanungo, 1994)
Envisioning (Bindl, Parker, Totterdell, & Hagger-Johnson, 2012)	In-Role Behavior (Williams & Anderson, 1991)	Individual Initiative (Moorman & Blakely (1995)
Equitable Organizational Practices (Nishii, 2013)	Perceived Organizational Support (Eisenberger, Armeli, Rexwinkel, Lynch, & Rhoades, 2001)	Organizational Identification (Mael & Ashforth, 1992)
Ethical Values Credits (Rodell & Lynch, 2016)	Values for Benevolence (Schwartz, 1994)	Prosocial Identity (Grant et al., 2008)
Evangelism Stigmas (Rodell & Lynch, 2016)	Machiavellianism (Jonason & Webster, 2010)	Values for Benevolence (Schwartz, 1994)
Exclusion Beliefs (Mitchell, Vogel, & Folger, 2015)	Machiavellianism (Jonason & Webster, 2010)	External Locus of Control (Levenson, 1981)
Expectation Reactions to Supervisor (Colquitt et al., 2015)	Satisfaction with Supervisor (Spector, 1985)	Trust in Supervisor (Brockner et al., 1997)
Extrinsic Career Goals (Seibert et al., 2013)	Promotion Satisfaction (Spector, 1985)	Values for Achievement (Schwartz, 1994)
Family Incivility (Lim & Tai, 2014)	Family Interference with Work (Gutek, Searle, & Klepa, 1991)	Subjective Stress (Motowidlo et al., 1986)
Family-to-Personal Conflict (Wilson & Baumann, 2015)	Values for Hedonism (Schwartz, 1994)	Subjective Stress (Motowidlo et al., 1986)
Gender Determinism (Tinsley, Howell, & Amanatullah, 2015)	Entity Theorist Beliefs (Chiu, Hong, & Dweck, 1997)	External Locus of Control (Levenson, 1981)
Global Organizational Commitment (Klein, Cooper, Molloy, & Swanson, 2014)	Perceived Organizational Support (Eisenberger et al., 2001)	Organizational Identification (Mael & Ashforth, 1992)
Idea Validation (Harrison & Wagner, 2016)	Task Proactivity (Griffin et al., 2007)	Individual Initiative (Moorman & Blakely, 1995)
Identity Strain (Kraimer, Shaffer, Harrison, & Ren, 2012)	Subjective Stress (Motowidlo et al., 1986)	Role Conflict (Peterson et al., 1995)
Inclusion in Decision Making (Nishii, 2013)	Participative Leadership (Arnold et al., 2000)	Supervisor Support (Oldham & Cummings, 1996)
Integration of Differences (Nishii, 2013)	Perceived Organizational Support (Eisenberger et al., 2001)	Organizational Identification (Mael & Ashforth, 1992)
Interactional Justice Clarity (Qin et al., 2015)	Satisfaction with Supervisor (Spector, 1985)	Trust in Supervisor (Brockner et al., 1997)
Interest/Enjoyment Sales Self-Efficacy (Gupta, Ganster, & Kepes, 2013)	Meaning (Spreitzer, 1995)	Job Satisfaction (Cammann et al., 1983)
International Employee Identity (Kraimer et al., 2012)	Workplace Status (Djurdjevic et al., 2017)	Values for Stimulation (Schwartz, 1994)
Interpersonal Fairness Monitoring (Long et al., 2011)	Feedback Seeking (Ashford, 1986)	Active Coping (Carver et al., 1989)
Intraorganizational Employee Navigation (Plouffe & Gregoire, 2011)	In-Role Behavior (Williams & Anderson, 1991)	Individual Initiative (Moorman & Blakely, 1995)
Intrinsic Career Goals (Seibert et al., 2013)	Learning Orientation (Vandewalle, 1997)	Values for Stimulation (Schwartz, 1994)
Knowledge Stock (Sung & Choi, 2012)	Workplace Status (Djurdjevic et al., 2017)	Generalized Self-Efficacy (Chen, Gully, & Eden, 2001)
Loss Orientation (Shepherd, Patzelt, & Wolfe, 2011)	Proactive Personality (Seibert, Crant, & Kraimer, 1999)	Active Coping (Carver et al., 1989)
Moral Credits (Lin, Ma, & Johnson, 2016)	Values for Benevolence (Schwartz, 1994)	Prosocial Identity (Grant et al., 2008)
Moral Identification (May, Chang, & Shao, 2015)	Organizational Identification (Mael & Ashforth, 1992)	Values for Benevolence (Schwartz, 1994)
Naiveté (Barasch, Levine, & Schweitzer, 2016)	Subjective Stress (Motowidlo et al., 1986)	Job-Related Affective Well-Being: Deactivated Negative Arousal (Van Katwyk et al., 2000)
Organizational Support for Development (Kraimer, Seibert, Wayne, Liden, & Bravo, 2011)	Perceived Organizational Support (Eisenberger et al., 2001)	Organizational Identification (Mael & Ashforth, 1992)

Table 2 (*continued*)

Focal Scale	Orbiting Scale 1	Orbiting Scale 2
Oscillation Orientation (Shepherd et al., 2011)	Proactive Personality (Seibert et al., 1999)	Active Coping (Carver et al., 1989)
Other Focus Credits (Rodell & Lynch, 2016)	Agreeableness (Donnellan et al., 2006)	Prosocial Identity (Grant et al., 2008)
Perceived Career Opportunity (Kraimer et al., 2011)	Perceived Organizational Support (Eisenberger et al., 2001)	Promotion Satisfaction (Spector, 1985)
Performance Monitoring (Guillaume, Knippenberg, & Brodbeck, 2014)	In-Role Behavior (Williams & Anderson, 1991)	Task Proactivity (Griffin et al., 2007)
Personal Life Motives (Leslie, Manchester, Park, & Mehng, 2012)	Family-Work Centrality (Carr, Boyar, & Gregory, 2008)	Family Interference with Work (Gutek et al., 1991)
Personal Recognition by Supervisor (Wang & Howell, 2010)	Supervisor Fairness (Colquitt et al., 2015)	Supervisor Consideration (Stogdill, 1963)
Personal-to-Family Conflict (Wilson & Baumann, 2015)	Family-Work Centrality (Carr et al., 2008)	Family Interference with Work (Gutek et al., 1991)
Personal-to-Work Conflict (Wilson & Baumann, 2015)	Family Interference with Work (Gutek et al., 1991)	Family-Work Centrality (Carr et al., 2008)
Planning (Bindl et al., 2012)	Active Coping (Carver et al., 1989)	In-Role Behavior (Williams & Anderson, 1991)
Procedural Fairness Monitoring (Long et al., 2011)	Feedback Seeking (Ashford, 1986)	Active Coping (Carver et al., 1989)
Productivity Motives (Leslie et al., 2012)	In-Role Behavior (Williams & Anderson, 1991)	Task Proactivity (Griffin et al., 2007)
Prohibitive Voice (Liang, Farh, & Farh, 2012)	In-Role Behavior (Williams & Anderson, 1991)	Civic Virtue (Podsakoff et al., 1990)
Promotive Voice (Liang et al., 2012)	In-Role Behavior (Williams & Anderson, 1991)	Civic Virtue (Podsakoff et al., 1990)
Propensity to Morally Disengage (Moore, Detert, Trevino, Baker, & Mayer, 2012)	Psychological Withdrawal (Lehman & Simpson, 1992)	Machiavellianism (Jonason & Webster, 2010)
Reflecting (Bindl et al., 2012)	Task Proactivity (Griffin et al., 2007)	In-Role Behavior (Williams & Anderson, 1991)
Relational Energy (Owens, Baker, Sumpter, & Cameron, 2016)	Workplace Friendships (Nielsen et al., 2000)	Coworker Satisfaction (Spector, 1985)
Reputation Maintenance Concerns (Baer et al., 2015)	Workplace Status (Djurdjevic et al., 2017)	Subjective Stress (Motowidlo et al., 1986)
Restoration Orientation (Shepherd et al., 2011)	Active Coping (Carver et al., 1989)	Proactive Personality (Seibert et al., 1999)
Self-/Other-Centeredness Supervisor Behaviors (Zhang et al., 2015)	Participative Leadership (Arnold et al., 2000)	Supervisor Consideration (Stogdill, 1963)
Self-Righteousness Stigmas (Rodell & Lynch, 2016)	Workplace Status (Djurdjevic et al., 2017)	Self-Esteem (Rosenberg, 1965)
Self-Verification Striving (Cable & Kay, 2012)	Proactive Personality (Seibert et al., 1999)	Values for Benevolence (Schwartz, 1994)
Sense of Community Credits (Rodell & Lynch, 2016)	Values for Benevolence (Schwartz, 1994)	Prosocial Identity (Grant et al., 2008)
Shame (Gonzalez-Gomez & Richter, 2015)	Job-Related Affective Well-Being: Deactivated Negative Arousal (Van Katwyk et al., 2000)	Subjective Stress (Motowidlo et al., 1986)
Skills/Ability Sales Self-Efficacy (Gupta et al., 2013)	Self-Esteem (Rosenberg, 1965)	Machiavellianism (Jonason & Webster, 2010)
Social Exchange Relationship (Colquitt, Baer, Long, & Halvorsen-Ganepola, 2014)	Satisfaction with Supervisor (Spector, 1985)	Trust in Supervisor (Brockner et al., 1997)
State Gratitude (Spence, Brown, Keeping, & Lian, 2014)	Agreeableness (Donnellan et al., 2006)	Job-Related Affective Well-Being: Deactivated Positive Arousal (Van Katwyk et al., 2000)
Subjective Social Class (Belmi & Neale, 2014)	Self-Esteem (Rosenberg, 1965)	Workplace Status (Djurdjevic et al., 2017)
Superficial Reactions to Supervisor (Colquitt et al., 2015)	Supervisor Fairness (Colquitt et al., 2015)	Job-Related Affective Well-Being: Deactivated Positive Arousal (Van Katwyk et al., 2000)
Supervisor Enforcing Work Requirements/Flexibility (Zhang et al., 2015)	Supervisor Support (Oldham & Cummings, 1996)	Supervisor Initiating Structure (Stogdill, 1963)
Supervisor Follower Development (Wang & Howell, 2010)	Supervisor Support (Oldham & Cummings, 1996)	Participative Leadership (Arnold et al., 2000)
Supervisor Intellectual Stimulation (Wang & Howell, 2010)	Achievement-Oriented Leadership (Indvik, 1985)	Participative Leadership (Arnold et al., 2000)

(*table continues*)

Table 2 (*continued*)

Focal Scale	Orbiting Scale 1	Orbiting Scale 2
Supervisor Procedural Justice Clarity (Qin et al., 2015)	Supervisor Initiating Structure (Stogdill, 1963)	Trust in Supervisor (Brockner et al., 1997)
Supervisor Solicitation of Voice (Fast, Burris, & Bartel, 2014)	Supervisor Consideration (Stogdill, 1963)	Participative Leadership (Arnold et al., 2000)
Supervisor Support for Recovery (Bennett, Gabriel, Calderwood, Dahling, & Trougakos, 2016)	Supervisor Consideration (Stogdill, 1963)	Participative Leadership (Arnold et al., 2000)
Supervisor Team Building (Wang & Howell, 2010)	Participative Leadership (Arnold et al., 2000)	Supervisor Vision Articulation (Conger & Kanungo, 1994)
Supervisor Vision Communication (Wang & Howell, 2010)	Achievement-Oriented Leadership (Indvik, 1985)	Supervisor Initiating Structure (Stogdill, 1963)
Supervisor-Triggered Newcomer Negative Affect (Nifadkar, Tsui, & Ashforth, 2012)	Subjective Stress (Motowidlo et al., 1986)	Job-Related Affective Well-Being: Activated Negative Arousal (Van Katwyk et al., 2000)
Supervisor-Triggered Newcomer Positive Affect (Nifadkar et al., 2012)	Satisfaction with Supervisor (Spector, 1985)	Job-Related Affective Well-Being: Activated Positive Arousal (Van Katwyk et al., 2000)
Supervisor's Organizational Embodiment (Eisenberger et al., 2010)	Supervisor Prototypicality (Ullrich et al., 2009)	Supervisor Vision Articulation (Conger & Kanungo, 1994)
Supportive Voice (Burris, 2012)	In-Role Behavior (Williams & Anderson, 1991)	Civic Virtue (Podsakoff et al., 1990)
Supportive Voice (Maynes & Podsakoff, 2014)	In-Role Behavior (Williams & Anderson, 1991)	Civic Virtue (Podsakoff et al., 1990)
Synchrony Preference (Leroy, Shipp, Blount, & Licht, 2015)	Agreeableness (Donnellan et al., 2006)	Openness (Donnellan et al., 2006)
Team Temporal Leadership (Mohammed & Nadkarni, 2011)	Achievement-Oriented Leadership (Indvik, 1985)	Supervisor Initiating Structure (Stogdill, 1963)
Tendency to Gossip (Erdogan, Bauer, & Walter, 2015)	Machiavellianism (Jonason & Webster, 2010)	Psychological Withdrawal (Lehman & Simpson, 1992)
Time Management Credits (Rodell & Lynch, 2016)	Conscientiousness (Donnellan et al., 2006)	Task Proactivity (Griffin et al., 2007)
Uniformity/Individualization Supervisor Behaviors (Zhang et al., 2015)	Supervisor Consideration (Stogdill, 1963)	Supervisor Fairness (Colquitt et al., 2015)
Victim Identity (Tepper, Mitchell, Haggard, Kwan, & Park, 2015)	Subjective Stress (Motowidlo et al., 1986)	Job-Related Affective Well-Being: Activated Negative Arousal (Van Katwyk et al., 2000)
Void Filling Stigmas (Rodell & Lynch, 2016)	Values for Stimulation (Schwartz, 1994)	Proactive Personality (Seibert et al., 1999)
Volunteering (Rodell, 2013)	Prosocial Identity (Grant et al., 2008)	Civic Virtue (Podsakoff et al., 1990)
Voracity (Rodell, 2013)	Learning Orientation (Vandewalle, 1997)	Values for Stimulation (Schwartz, 1994)
Wanderlust (Rodell, 2013)	Proactive Personality (Seibert et al., 1999)	Job-Related Affective Well-Being: Deactivated Negative Arousal (Van Katwyk et al., 2000)
Work-to-Personal Conflict (Wilson & Baumann, 2015)	Family Interference with Work (Gutek et al., 1991)	Subjective Stress (Motowidlo et al., 1986)

person enjoys stimulating and exciting activities, even if they involve some danger" (p. 736, emphasis in original). The judges were then shown the 35 items on a separate sheet of paper, again presented in random order. For example, one item was presented as "______ I take dares just for fun" (p. 735). Judges were then asked to fill in the blank with either energetic, impulse purchaser, thrill-seeking, avoids complexity, or unreflective, depending on which construct they believed was represented by the item. Anderson and Gerbing (1991) did not present full results for either the 35 items or the five scales, though they did show that p_{sa} and c_{sv} values predicted items' factor loadings in a subsequent confirmatory factor analysis.

Hinkin and Tracey (1999)

Hinkin and Tracey (1999) introduced their approach by noting that most scholars fail to use any content validation approach, or to document the approach that they actually use. Building on an earlier technique by Hinkin and Tracey's (1999), Schriesheim, Powers, Scandura, Gardiner, and Lankau (1993) approach contains many similarities to Anderson and Gerbing's (1991). Construct definitions again play a central role, with judges comparing scale items to definitions. In addition, naïve judges are again preferred, with Hinkin and Tracey (1999) writing that the only requirement for their task is "sufficient intellectual ability to rate the correspondence between items and definitions of various theoretical constructs, and the lack of any pertinent biases" (p. 179).

One key difference with Hinkin and Tracey's (1999) approach is that Anderson and Gerbing's (1991) sorting process is replaced by a Likert-style rating process. In their initial demonstration, judges rated how well scale items corresponded to the construct's definition using 1 (*not at all*) to 5 (*completely*) anchors. The resulting average rating provides a straightforward index of definitional correspondence—referred to by the authors as "content adequacy." For example, knowing that a scale's items have an average definitional correspondence of 4.20 provides straightforward information on content adequacy. Of course, other scholars may decide to use a 7-point scale rather than a 5-point scale—with a 4.20 becoming less supportive in that circumstance. We therefore propose an index termed *htc* (for *Hinkin Tracey correspondence*) that divides the average definitional correspondence rating across a scale's items by *a*, the number of anchors. The *htc* statistic would therefore take on a perfect value of 1 when all judges selected the maximum anchor for all scale items.

$$htc = \text{average definitional correspondence rating}/a$$

Importantly, the judges in Hinkin and Tracey's (1999) approach also rate how well scale items correspond to orbiting constructs using the same Likert-style rating. Each scale item therefore winds up with multiple signed differences—differences between the rating on its intended construct and the ratings on unintended constructs. Hinkin and Tracey (1999) described an ANOVA procedure for examining those differences, but the differences themselves provide a benchmark for definitional distinctiveness. For example, consider a scale that is validated using a 7-point scale, including two orbiting constructs. Knowing that the scale has an average definitional correspondence rating of 5.80 on its construct, and an average definitional correspondence rating of 3.60 and 4.20 on the two orbiting constructs, is informative. We therefore propose an index called *htd* (for *Hinkin Tracey distinctiveness*) that averages together all of those signed differences before dividing by $a - 1$, where *a* is again the number of anchors. Subtracting one from the number of anchors gives this index a range from negative 1.00 to 1.00, no matter the number of anchors used. In the example above, *htd* would be the average of (5.80 minus 3.60) and (5.80 minus 4.20), with that average then divided by six (seven scale anchors minus one). That figure would be 1.90/6, which equals .32.

$$htd = \text{Average of all (Intended Correspondence Rating} \\ - \text{Orbiting Correspondence Rating})/(a - 1)$$

The *htd* statistic would have a positive value when items received higher ratings on the intended construct than on the orbiting ones and a negative value when items received lower ratings on the intended construct than on the orbiting ones. Thus, as with the other three statistics benchmarked here, "higher is better." A theoretical upper bound would be an average definitional correspondence rating of 7.00 on a scale's intended construct and an average definitional correspondence rating of 1.00 on each of the orbiting constructs. With two orbiting constructs, *htd* would be the average of (7.00 minus 1.00) and (7.00 minus 1.00), with that average then divided by six (seven scale anchors minus one). That figure would be 6.00/6, which equals 1.00. A theoretical lower bound, by extension, would be when that rating pattern was flipped, with an average definitional correspondence rating of 1.00 on a scale's intended construct and an average definitional correspondence rating of 7.00 on each of the orbiting constructs. There *htd* would be the average of (1.00 minus 7.00) and (1.00 minus 7.00), with that average then divided by six. That figure would be negative 6.00, or negative 1.00.

Hinkin and Tracey (1999) demonstrated their approach, in part, using the four facets of transformational leadership from the Multifactor Leadership Questionnaire (Bass & Avolio, 1990): idealized influence (10 items), inspirational motivation (10 items), intellectual stimulation (10 items), and individualized consideration (nine items). A sample of 57 graduate business students served as the judges. The judges were given a four-page questionnaire, with each page having the definition of a transformational leadership facet at the top and all 39 items beneath. The four definitions were presented in random order, as were the 39 items. The judges rated the correspondence between items and definitions using the 5-point scale given previously. Although they did not present their results using our formulas, the four scales would have earned average *htc*'s ranging from .81 to .89 and average *htd*'s ranging from .11 to .24.

Applications of the Approaches

The content validation approaches introduced by Anderson and Gerbing (1991) and Hinkin and Tracey (1999) both seem to offer useful information for gauging the validity of inferences made with scale measures. The analyses used to gauge definitional correspondence and definitional distinctiveness are straightforward and speak directly to the degree to which scale items adequately sample the universe of content associated with a construct. Despite that practicality and relevance, however, applications of both approaches remain rare in top journals. Indeed, of the 56 articles that introduced new scales in *Journal of Applied Psychology*, *Academy of Management Journal*, *Personnel Psychology*, and *Organizational Behavior and Human Decision Processes* from 2010 to 2016, 33 reported no content validation of any kind. In contrast, only one of the 56 articles omitted reliability information, with only seven failing to report examinations of factor structure. Those 33 articles contributed a total of 61 scales to the literature, meaning that the literature now possesses 61 scales that may or may not adequately reflect their intended constructs of interest.

Of the 56 articles that introduced new scales, 13 reported a content validation that included some elements of Anderson and Gerbing's (1991) or Hinkin and Tracey's (1999) approaches. For example, some sorted items into multiple construct definitions (Erdogan, Bauer, & Walter, 2015; Fast, Burris, & Bartel, 2014; Gonzalez-Gomez & Richter, 2015; Nifadkar, Tsui, & Ashforth, 2012; Nishii, 2013; Wilson & Baumann, 2015; Zhang, Waldman, Han, & Li, 2015) while others rated the degree to which items matched construct definitions (Dobrow & Tosti-Kharas, 2011; Hannah, Jennings, Bluhm, Peng, & Schaubroeck, 2014; Klein, Cooper, Molloy, & Swanson, 2014; Leroy, Shipp, Blount, & Licht, 2015). However, the descriptions of many efforts lacked the details needed to understand the results (Fast et al., 2014; Hannah et al., 2014; Leroy et al., 2015; Liang, Farh, & Farh, 2012). In cases where adequate details were provided, the criteria for retaining items varied widely and seemed arbitrary (Dobrow & Tosti-Kharas, 2011; Erdogan et al., 2015; Gonzalez-Gomez & Richter, 2015; Klein et al., 2014; Nishii, 2013; Wilson & Baumann, 2015).

In terms of judges, some included naïve judges in their samples (Nifadkar et al., 2012; Nishii, 2013; Wilson & Baumann, 2015), some used both faculty and doctoral students (Fast et al., 2014; Klein et al., 2014; Wilson & Baumann, 2015), while others used exclusively doctoral students (Erdogan et al., 2015; Leroy et al., 2015; Liang et al., 2012; Zhang et al., 2015). The number of judges was often smaller than those used in Anderson and Gerbing (1991) or Hinkin and Tracey (1999), with a median of 10 and a mode of only three.

Of the 56 articles that introduced new scales, three made an explicit reference to Anderson and Gerbing's (1991) approach (Harrison & Wagner, 2016; Shepherd, Patzelt, & Wolfe, 2011; Spence, Brown, Keeping, & Lian, 2014). One of those used nine expert judges (six doctoral students and three professors) and did not report either p_{sa} or c_{sv} (Shepherd et al., 2011). Another used seven expert judges (doctoral students) while reporting a c_{sv} of .89 but no p_{sa} (Harrison & Wagner, 2016). Only the third used the naïve judges recommended by Anderson and Gerbing (1991), employing 24 employed adults recruited through snowball sampling (Spence et al., 2014). That application noted that items were retained if they possessed a p_{sa} and c_{sv} of .75—a hurdle taken from Hinkin's (1998) discussion of his doctoral dissertation (Hinkin, 1985).

That leaves seven articles that made an explicit reference to Hinkin and Tracey's (1999) approach (Baer et al., 2015; Bolino, Hsiung, Harvey, & LePine, 2015; Colquitt, Baer, Long, & Halvorsen-Ganepola, 2014; Long, Baer, Colquitt, Outlaw, & Dhensa-Kahlon, 2015; Maynes & Podsakoff, 2014; Rodell, 2013; Rodell & Lynch, 2016). All used naïve judges, with all using undergraduates as Hinkin and Tracey (1999) had done. The sample sizes varied widely, from a low of 47 (Maynes & Podsakoff, 2014) to a high of 782 (Rodell, 2013). Although most applications employed a seven-point scale, the wording of the anchors varied from article to article (while only providing wording for the endpoints). Such variation and ambiguity matters because the anchor wording can impact the average rating an item earns. For example, the likelihood of an item earning a "7" may differ when the specific anchor is *completely captured by the conceptual definition* (Maynes & Podsakoff, 2014, p. 94) versus *extremely good match to the definition* (Long et al., 2015, p. 471). For the six articles who reported correspondence levels, their average *htc* would have ranged from .74 (Long et al., 2015) to .98 (Maynes & Podsakoff, 2014). Interestingly, only three articles examined definitional distinctiveness (Bolino et al., 2015; Colquitt et al., 2014; Maynes & Podsakoff, 2014). The average *htd* results for those articles would have ranged from .06 to .97—likely as a function of how similar the focal construct is to the orbiting constructs.

Generating Our Evaluation Criteria

When reviews point out that the correspondence between constructs and measures is often tenuous in industrial/organizational psychology and organizational behavior (Aguinis & Edwards, 2014; see also Aguinis & Vandenberg, 2014 and Stone-Romero, 1994), it is easy to argue that content validation should become more frequent. We are, however, sympathetic to authors who see some ambiguity in Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches. Indeed, that ambiguity becomes especially acute when applications of the approaches vary type of judges, number of judges, number of anchors, wording of anchors, and so forth. Moreover, the limited applications of those techniques—together with such variation—makes it difficult to put one's results in a larger context.

By subjecting the 112 scales introduced in *Journal of Applied Psychology*, *Academy of Management Journal*, *Personnel Psychology*, and *Organizational Behavior and Human Decision Processes* to Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approach, we hope to provide clear evaluation criteria for p_{sv}, c_{sv}, *htc*, and *htd*. We document our application of those approaches in a detailed, transparent, and replicable fashion so that future scholars can apply them in a way that yields "apples to apples" comparisons. Importantly, we wanted to create our guidelines using judges who are faithful to Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) descriptions. Our judges therefore needed to be "representative of the main study sample and population of interest" (Anderson & Gerbing, 1991, p. 734) with "sufficient intellectual ability to rate the correspondence between items and definitions of various theoretical constructs" (Hinkin & Tracey, 1999, p. 179).

We drew our judges from Amazon's Mechanical Turk (MTurk; Paolacci & Chandler, 2014; Sheehan & Pittman, 2016), with the stipulation that the participants be employed and that they reside in the United States. The employment stipulation speaks to the representativeness requirement from Anderson and Gerbing (1991) insofar as most studies in industrial/organizational psychology and organizational behavior are focused on working adults. The United States stipulation speaks to the skill requirement from Hinkin and Tracey (1999) insofar as the definitions and scale items in our sample were published in English. Judges living in an English-speaking area should be more attuned to close synonyms between definitions and items, or to intended and unintended nuances in them. As a source of naïve judges, MTurk offers two potential advantages relative to undergraduate subject pools (Paolacci & Chandler, 2014; Sheehan & Pittman, 2016). First, it is widely available, including to scholars at small colleges and universities with limited class sizes. Second, it can be sampled with specific parameters that—if replicated—can create the potential for more "apples to apples" comparisons, relative to subject pools from very different colleges and universities.

In generating our evaluation criteria, it was necessary to choose orbiting constructs in order to calculate statistics relevant to item distinctiveness. The orbiting constructs we chose are shown in Table 2. Such choices are pivotal to the content validation process given the impact they can have on our statistics. It is therefore important to understand the specific guidelines we used to make those choices, which can also provide guidance for other scholars. We provide four different guidelines below.

First, we chose venerable constructs that scholars might often consider when performing discriminant validation analyses with new scales. Some of those venerable constructs included the Big Five dimensions of personality, self-esteem, job satisfaction, subjective stress, coworker support, participative leadership, and in-role performance. The use of venerable constructs allows new scales to be judged against constructs whose definitions are well understood and whose scales have been used quite frequently in the literature.

Second, we focused on constructs that were at the same stage of "causal flow"—not typically being viewed as either antecedents or consequences of the focal construct. For example, Kraimer et al.'s (2012) identity strain was paired with subjective stress and role conflict because both are state variables that would be viewed as correlates. In contrast, we would not have focused on an antecedent like neuroticism or a consequence like somatic symptoms. Forcing orbiting constructs to be at the same stage of causal flow provides a higher bar for definitional distinctiveness, given that antecedent and consequence variables should naturally be more conceptually and empirically distinct. Moreover, constructs at the same stage of causal flow are more likely to be at issue in debates about construct proliferation (Shaffer, DeGeest, & Li, 2016)—an issue we return to in the Discussion section.

Third, we avoided choices that had a "part-whole" relationship with the focal construct, where one was a subfacet of—or more specific instance of—the other. For example, Liang, Farh, and Farh's (2012) promotive voice was paired with in-role behavior and civic virtue rather than a measure of voice in general (Van Dyne & LePine, 1998). A construct that is a more specific instance of a broader construct would not be expected to be definitionally distinct from the broader construct. It might be definitionally distinct from *a different subfacet of* the broader construct, of course, though that question is more relevant to our second guideline.

Fourth, we attempted to choose orbiting constructs that utilized the same referent as the focal construct. For example, Colbert, Bono, and Purvanova's (2016) coworker friendship was paired with coworker support and coworker satisfaction. The definitions of two constructs will seem artificially distinct if they possess different referents, even if the conceptual core of both is identical. Attempting to utilize the same referent again provides a higher bar for definitional distinctiveness.

The few results that have been published for definitional distinctiveness illustrate how impactful the choice of orbiting constructs is for one's results. Hinkin and Tracey's (1999) use of multiple transformational leadership facets resulted in average *htd*'s as low as .11. Colquitt, Baer, Long, and Halvorsen-Ganepola's (2014) use of multiple conceptualizations of social exchange for their social exchange scale resulted in an average *htd* of .06. In contrast, Bolino, Hsiung, Harvey, and LePine (2015) used more distinct constructs for their content validation of citizenship fatigue—citizenship pressure and burnout—earning an average *htd* of .37. It may therefore be that—all else equal—definitional distinctiveness results will be lower when a focal construct is more similar to the orbiting constructs.

To explore this possibility—and to build it into our evaluation criteria—we gathered substantive data on the constructs shown in Table 2. Those data allowed us to see if, for example, Dobrow and Tosti-Kharas's (2011) calling scale was more highly correlated with scales for job satisfaction and job-related affective well-being than Belmi and Pfeffer's (2016) desire for power scale was with scales for extraversion and Machiavellianism. As another example, we could see whether Owens, Baker, Sumpter, and Cameron's (2016) relational energy scale was more highly correlated with scales for workplace friendships and coworker satisfaction than Rodell's (2013) volunteering scale was with scales for prosocial identity and civic virtue. To the extent that such differences manifested in a substantive data collection, we could explore whether they translated to differences in definitional distinctiveness results. If they did, we could supplement our evaluation criteria with more nuanced guidelines that took into account the average correlation between the focal scale and the orbiting scales. Given that our criteria generation utilized MTurk, we drew on MTurk again for this "correlation-based norming." Note that this substantive data collection injected an additional criterion for choosing the orbiting constructs, insofar as we wanted to reuse choices where possible to keep the substantive data collection more manageable.

Method

Generating Evaluation Criteria

Given that our evaluation criteria were based on 112 scales, it was obviously impossible for any one participant to complete Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approach for all of them. Our benchmarking was therefore based on multiple "subsamples" rather than one large sample. We created subsamples that would not tax our respondents and that seemed commensurate with typical content validation efforts. Those criteria resulted in 48 subsamples that included five different rows from Table 2—chosen at random—with each of the rows subjected to either Anderson and Gerbing's (1991) approach or Hinkin and Tracey's (1999) approach. For example, one subsample subjected global organizational commitment (Klein et al., 2014), propensity to morally disengage (Moore, Detert, Trevino, Baker, & Mayer, 2012), victim identity (Tepper, Mitchell, Haggard, Kwan, & Park, 2015), communicating high expectations by supervisor (Wang & Howell, 2010), and evangelism stigmata (Rodell & Lynch, 2016) to Anderson and Gerbing's (1991) approach, with each scale compared to the orbiting scales shown in its row. Another subsample did the same for Hinkin and Tracey's (1999) approach. Leaving aside instructions, demographics, and so forth, these decisions resulted in surveys that were an average of 159 items in length, presented across an average of 10 pages within Qualtrics (IRB Protocol Number: STUDY00002856; Title: Content Validity of Management Constructs; Institutional Review Board at the University of Georgia).

Having divided our 112 scales into 48 subsamples, it was necessary to choose an appropriate size for those subsamples. We wanted that choice to be reasonably consistent with applications of the content validation approaches, and also large enough to make our evaluation criteria sufficiently precise. Taking Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) own applications of their approaches, along with the 10 applications of them reviewed previously, results in a median sample size of 70 judges. We felt that expanding that guideline to a minimum sample size of 100 would result in means for our evaluation criteria that were sufficiently precise. The standard error of the mean is the standard deviation divided by the square root of the sample size. Taking into account the standard deviations that were eventually derived for our statistics, the standard errors of the means with $N = 100$ would range from .006 to .029 with an average of .016.

Our survey length and sample size decisions resulted in the recruitment of 48 MTurk subsamples that initially averaged 130 participants. In addition to the stipulation that participants be employed and residing in the United States, we utilized two data quality requirements: participants needed to have completed 50 tasks through the system and needed to have earned at least a 90% approval on those tasks (see Sheehan & Pittman, 2016, for a discussion of such issues). In addition, to prevent nonindependence, MTurkers were not allowed to be part of more than one subsample. We paid MTurkers $2 for their participation. Thus, our sample for generating evaluation criteria included a total of 6,240 participants who earned a total payout of $12,480.

Participants who utilized Anderson and Gerbing's (1991) approach did so using the instructions in Appendix A; those who utilized Hinkin and Tracey's (1999) approach used the instructions in Appendix B. Both instructions were designed to be detailed, transparent, and replicable. Notable elements include a discussion of content validation studies in general, an illustration of how to present a construct label alongside its definition, a discussion of how reverse-worded items can still correspond to a construct definition, and the inclusion of a quiz to verify understanding. In the case of Anderson and Gerbing (1991), participants were shown how they would "drag and drop" items into the box associated with the most relevant construct and definition. In the case of Hinkin and Tracey (1999), participants were shown the response scale they would utilize in rating the items, ranging from 1 = *Item does an **EXTREMELY BAD** job of measuring the **bolded** concept provided above* to 7 = *Item does an **EXTREMELY GOOD** job of measuring the **bolded** concept provided above*.

Upon beginning the survey, participants encountered one of the five focal constructs—together with the two orbiting constructs—chosen at random. For example, consider a participant in the subsample described previously who utilized Hinkin and Tracey's (1999) approach. The participant might begin her survey with the propensity to morally disengage construct (Moore et al., 2012). That construct label and definition might appear first, followed by the items for propensity to morally disengage (Moore et al., 2012), psychological withdrawal (Lehman & Simpson, 1992), and Machiavellianism (Jonason & Webster, 2010)—all interspersed in a random order. If that participant instead utilized Anderson and Gerbing's (1991) approach, boxes with all three construct labels and definitions would appear first. The items for all three scales would then follow—again interspersed in random order. The survey would then continue with some randomized combination of global organizational commitment (Klein et al., 2014), victim identity (Tepper et al., 2015), communicating high expectations by supervisor (Wang & Howell, 2010), and evangelism stigmata (Rodell & Lynch, 2016).

We conducted several checks to maximize data quality. For example, we monitored the time spent on each page of the survey and the overall time taken for the survey as a whole. We also included three different careless respondent checks where participants were instructed to click a particular rating or drag the item into a particular box (Meade & Craig, 2012). These checks, together with listwise deletion of missing data, resulted in a total sample of 5,150 participants—83% of the total participants recruited. That level of attrition is similar to recent *Journal of Applied Psychology* studies with similar designs (e.g., Dang, Umphress, & Mitchell, 2017; Hideg & Ferris, 2017; Landis, Kilduff, Menges, & Kilduff, 2018; Van Dijke, De Cremer, Langendijk, & Anderson, 2018). The average sample size for our 48 subsamples was 108. That level exceeds the minimum sample size of 100 that was our goal, and the median sample size of 70 from Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) own applications of their approaches. The participants were 55% female, were an average of 36.3 years old (*SD* = 10.78), and were 77% Caucasian, 8% African American, 6% Asian, 6% Hispanic, and 3% Other.

Correlation-Based Norming

The data collection for our correlation-based norming required similar decisions to our generation of evaluation criteria. Given that we reused orbiting constructs where possible, Table 2 includes a total of 168 unique scales. It was therefore impossible for any one participant to provide substantive ratings on all 168 scales. As a result, we divided the constructs into seven different surveys. Leaving aside instructions, demographics, and so forth, that decision resulted in surveys that were an average of 163 items in length, presented across an average of nine pages within Qualtrics.

Our sample size decision making followed a similar logic to our evaluation criteria, as we wanted to be reasonably consistent with most substantive studies in the literatures embedded in Table 2 while being large enough to make our correlations sufficiently precise. We felt that a minimum sample size of 300 would achieve those goals. The standard error of (Fisher's z transformation of) the correlation is one divided by the square root of the sample size minus three. The standard error would therefore be .06 with N = 300.

Our survey length and sample size decisions resulted in the recruitment of seven MTurk subsamples that averaged 334 participants. We used the same recruitment requirements that were employed to generate our evaluation criteria, and again stipulated that participants were not allowed to be part of more than one subsample. We again paid MTurkers $2 for their participation. Thus, our correlation-based norming sample included a total of 2,335 participants who earned a total payout of $4,670. We employed the same checks for maximizing data quality, with those checks and listwise deletion of missing data resulting in a total sample of 2,119 participants—91% of the total participants recruited. The average sample size for our seven subsamples was 303. That level exceeds the minimum sample size of 300 that was our goal. The participants were 51% male, were an average of 37.3-years-old (*SD* = 11.02), and were 77% Caucasian, 8% African American, 7% Asian, 6% Hispanic, and 2% Other. The participants responded to the scale items on their survey in the same way respondents in any substantive study would—by indicating agreement on a 5-point scale with anchors of 1 = *strongly disagree* to 5 = *strongly agree*. That is, the participants whose surveys included bottom-line mentality (Greenbaum, Mawritz, & Eissa, 2012), supervisor initiating structure (Stogdill, 1963), and achievement-oriented leadership (Indvik, 1985) indicated the degree to which they agreed that their boss engaged in those sorts of behaviors, just as they would have in a substantive study.

Results

Distributional Properties of Statistics

To summarize our statistics for the 112 scales in our review, we calculated the average p_{sa}, c_{sv}, *htc*, and *htd* across the multiple items. The average is commonly used in other realms of scale validation, as when reporting the average factor loading for a scale's items in the measurement model. Thus, our use of the p_{sa}, c_{sv}, *htc*, and *htd* terms moving forward reflects averages across a scale's items. Online supplement Table S1 provides the p_{sa}, c_{sv}, *htc*, and *htd* for all 112 scales. The table also provides the number of items in the scale, its coefficient alpha in our correlation-based norming data, and the orbiting scales that we utilized. Figures 1, 2, 3, and 4 provide frequencies for the 112 values for those four statistics at the scale level. As would be expected, most distributions are negatively skewed—bumping up against perfect values on the right side and being pulled down by lower values on the left side.

Table 3 summarizes the distributional properties of the four statistics at the scale level. Beginning with Anderson and Gerbing's (1991) statistics, the p_{sa} values ranged from .24 to .98, with a mean of .79, a median of .82, and a standard deviation of .15. Recall that p_{sa} takes on a value of 0 when no judges classify an item correctly and a value of 1 when all judges classify an item correctly. The c_{sv} values ranged from negative .52 to .95, with a mean of .57, a median of .63, and a standard deviation of .29. Recall that c_{sv} takes on a value of negative 1 when no judges classify an item correctly and all do so incorrectly and a value of 1 when all judges classify an item correctly and none do so incorrectly.

Turning to the Hinkin and Tracey (1999) statistics introduced here, the *htc* values ranged from .60 to .96, with a mean of .87, a median of .88, and a standard deviation of .06. Recall that *htc* takes on a value of 1 when all judges select the maximum anchor for all scale items. Finally, the *htd* values ranged from negative .04 to .64, with a mean of .27, a median of .26, and a standard deviation of .14. Recall that *htd* has a positive value when items receive higher ratings on the intended construct than on the orbiting ones and a negative value when items receive lower ratings on the intended construct than on the orbiting ones. Theoretical lower and upper bounds for *htd* are negative 1.00 and 1.00, respectively.

Table 4 illustrates correlations between the properties of the 112 scales and their p_{sa}, c_{sv}, *htc*, and *htd* values. Longer scales tended to have lower *htc* values, perhaps because there is a limit to how many ways an item can capture a definition. Scales that utilized reverse-worded items had lower levels of all four statistics. The use of such items has been a subject of some debate (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Schmitt & Stults, 1985; Schriesheim & Eisenbach, 1995). Anderson and Gerbing (1991) included reverse-worded items in their demonstration of their approach—with no discussion of differences in results—whereas Hinkin and Tracey (1999) used only regular items. Although only five scales included reverse-worded items, their statistics were different enough to yield significant negative correlations with a dummy variable indicating the presence of such items. The percentage of such items in a given scale yielded weaker relationships, however, being correlated only with *htc*. As would be expected given the reverse-worded effects, the coefficient alpha for a scale (taken from our correlation-based norming data) was positively related to the scale's p_{sa}, c_{sv}, *htc*, and *htd*.

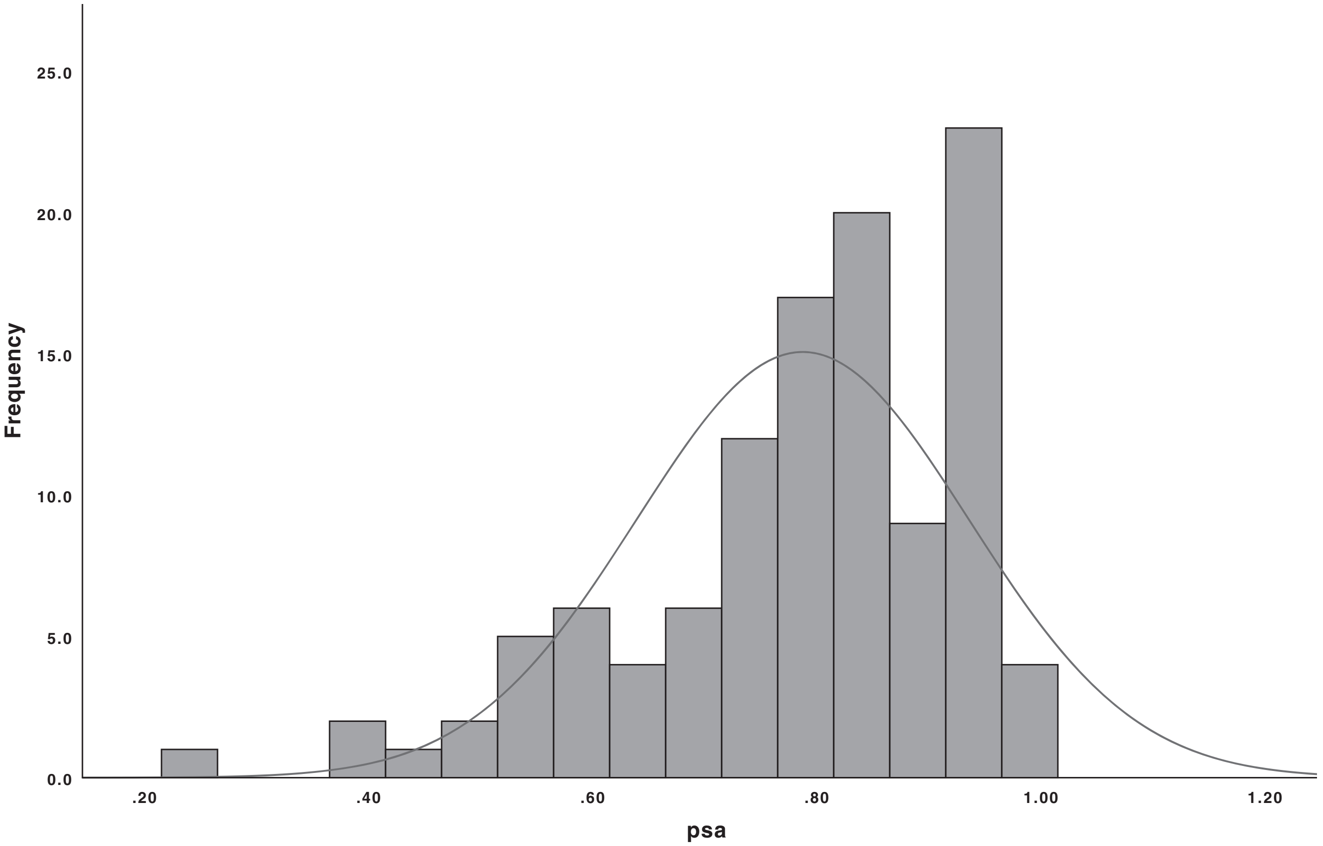


Figure 1. Distribution of p_{sa} for the 112 scales used to generate our evaluation criteria.

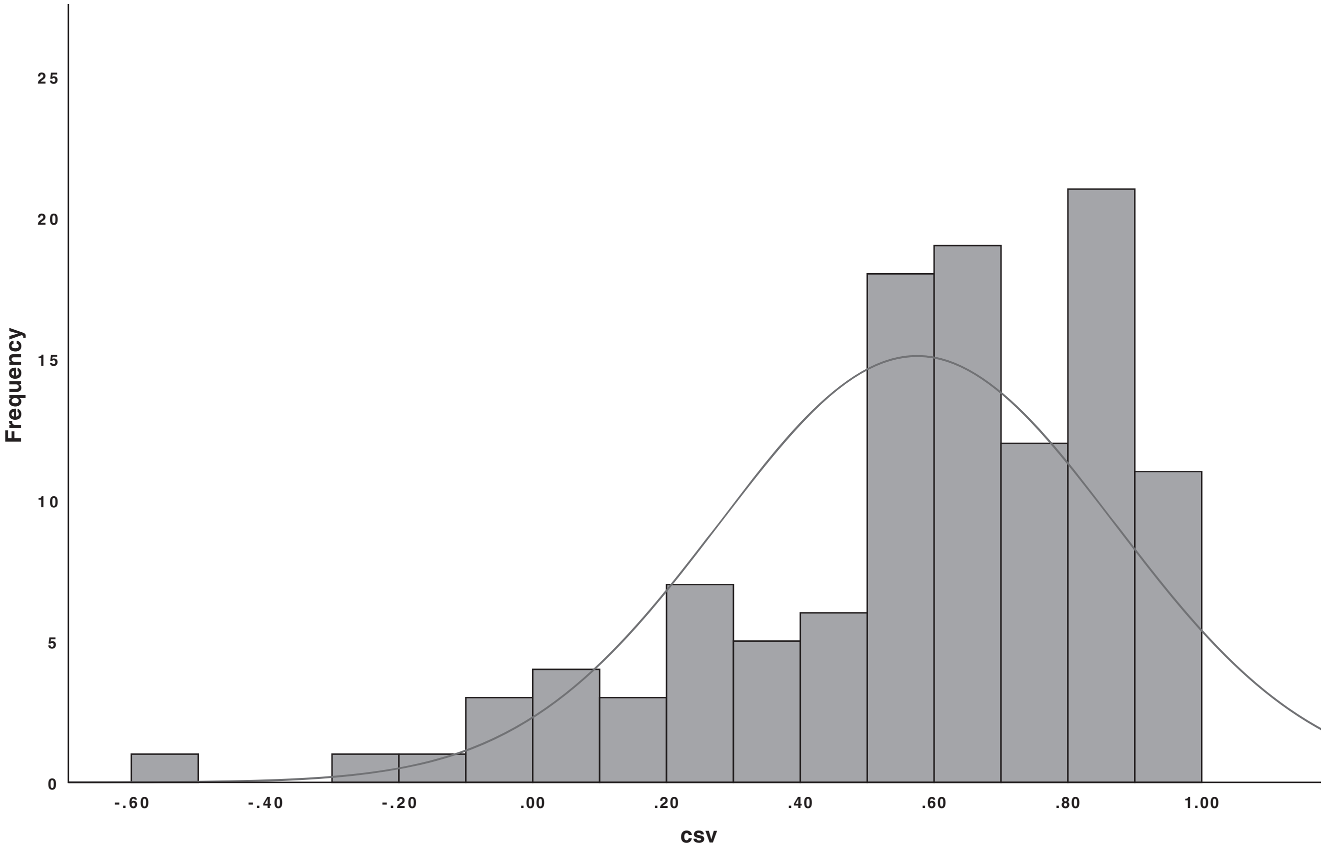


Figure 2. Distribution of c_{sv} for the 112 scales used to generate our evaluation criteria.

Table 4 also reveals that there was no relationship between whether an article employed some sort of content validation and the p_{sa}, c_{sv}, *htc*, and *htd* that resulted from the scale. Recall that 23 of the 56 articles reported some validation approach that had at least some elements of Anderson and Gerbing (1991) or Hinkin and Tracey (1999), even if no reference was made to either. Similarly, there was no relationship between whether an article employed Anderson and Gerbing's (1991) approach and the p_{sa}, c_{sv}, *htc*, and *htd* that resulted from the scale. It should be noted, however, that only three articles employed their approach, with only one using the naïve judges that were recommended (Spence et al., 2014). In contrast, the use of Hinkin and Tracey's (1999) approach was associated with higher levels of p_{sa}, c_{sv}, and *htc*. Recall that seven of the 56 articles utilized Hinkin and Tracey's (1999) approach. Finally, Table 4 reveals that higher average correlations with a scale's orbiting constructs were associated with lower levels of p_{sa}, c_{sv}, and *htc*. Those results point to the need for correlation-based evaluation criteria, as will be described in a subsequent section.

Overall Evaluation Criteria

We created guidelines for the definitional correspondence and definitional distinctiveness statistics using LeBreton and Senter's (2008) discussion of interrater agreement as a guide. Their Table 3 provides ranges of agreement values that correspond to interpretations ranging from *very strong agreement* to *lack of agreement*. To create that same sort of structure, we created quintiles for all four statistics to split their distributions into five percentile ranges. The percentile ranges created by those quintiles are shown in the top panel of Table 5. In lieu of using the lower bounds and upper bounds for each statistic from Table 3, Table 5 uses "and below" and "and above" for the outer ranges.

Assume, for example, that a scholar used Anderson and Gerbing's (1991) approach, with the results revealing a p_{sa} of .93. The evaluation criteria in the top panel of Table 5 would label that value as "very strong," given that it is in the 80th to 99th percentile of the 112 p_{sa} values earned by new scales introduced in *Journal of Applied Psychology*, *Academy of Management Journal*, *Personnel Psychology*, *and Organizational Behavior and Human Decision Processes* from 2010–2016. Alternatively, assume that a scholar used Hinkin and Tracey's (1999) approach, with the results revealing an *htc* of .61. The evaluation criteria would label that value as "weak," given that it is in the 20th to 39th percentile of the 112 *htc* values earned by new scales introduced in those journals in that time window.

Correlation-Based Evaluation Criteria

The evaluation criteria in the top panel of Table 5 already provide something largely missing from the literature on content validation—some way to evaluate the numbers derived from An-

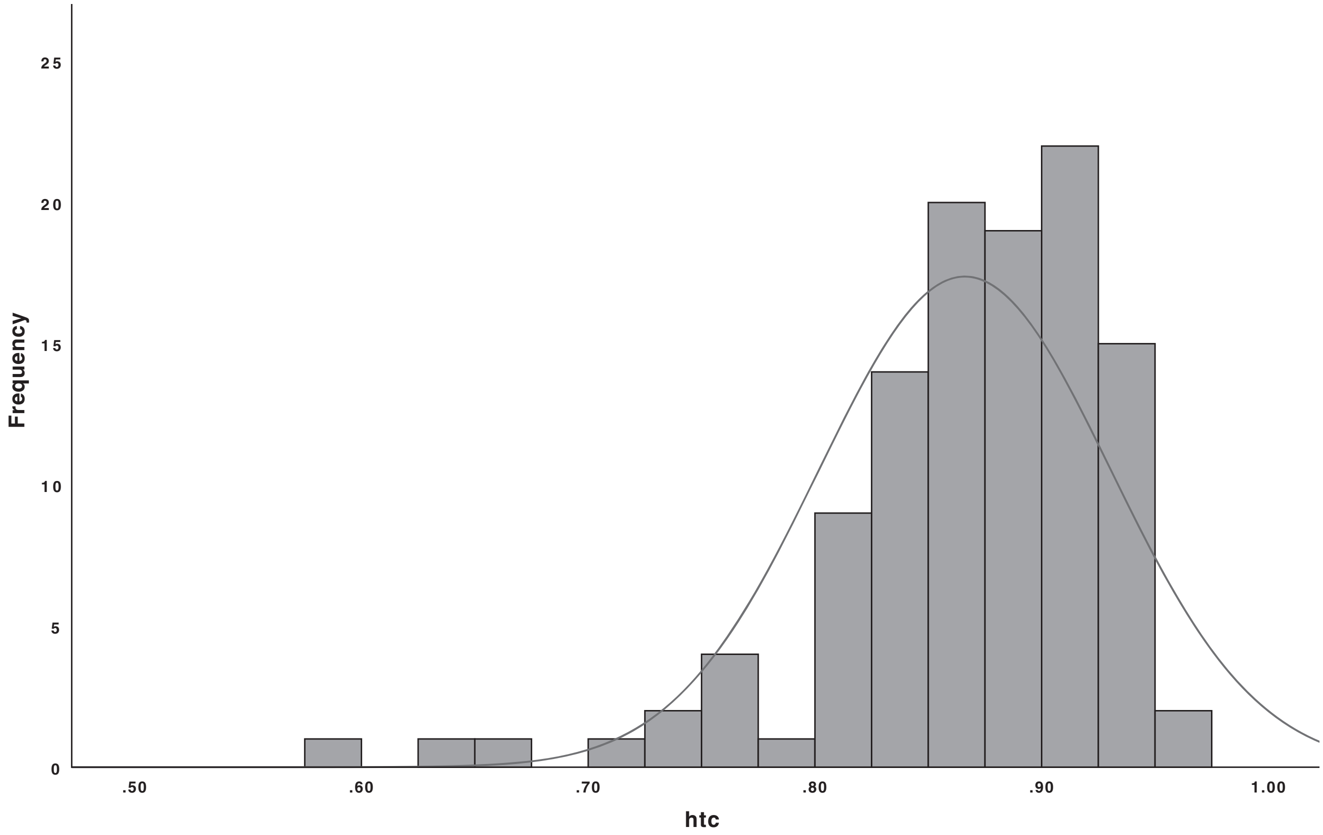


Figure 3. Distribution of *htc* for the 112 scales used to generate our evaluation criteria.

derson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches. That said, Table 4 revealed that p_{sa}, c_{sv}, and *htc* results may depend on the choice of orbiting constructs. If a focal construct is more similar to an orbiting construct—conceptually or empirically—then statistics that ask the participant to simultaneously consider the two may yield lower values. We therefore created correlation-based evaluation criteria that could take this nuance into account. The average correlation between focal scales and their two orbiting scales ranged from .02 to .84 across our 112 scales, with a mean of .43, a median of .41, and a standard deviation of .19. We created tertiles of those 112 average correlation values to split that .02 to .84 into three percentile ranges: 0th to 32nd percentile (.02 to .34), 33rd to 66th percentile (.35 to .50), and 67th to 99th percentile (.51 to .84). As we did in the top panel of Table 5, we replaced the lower bound and upper bound of that range with "or below" and "or above." Finally, for the scales that fell within each of those ranges, we recreated the quintiles for our four statistics to provide the same interpretational categories. The resulting structure is shown in the other three panels of Table 5.

Assume, for example, that a scholar used Anderson and Gerbing's (1991) approach, with the results revealing a c_{sv} of .70. Our correlation-based evaluation criteria would put that value in the "moderate" category if the average correlation between the focal scale and its orbiting scales was .34 or below, in the "strong" category if that average correlation was .35 to .50, and in the "very strong" category if that average correlation was .51 or above. The more similar the focal scale becomes to the orbiting scales, the more "impressive" a given level of definitional distinctiveness becomes.

Discussion

Consider for a moment the sheer magnitude of Table 2. A total of 112 new scales were introduced in four prominent industrial/organizational psychology and organizational behavior outlets from 2010 to 2016. From avoidant leader behaviors to work-to-personal conflict, these scales will wind up joining the research agendas of many scholars. Now consider that there are 79 journals in Web of Science's "psychology, applied" category. It may be that new scales are introduced less frequently in some of those other journals, if their mission statements do not emphasize important contributions to the same degree as the outlets examined here. Still, a conservative extrapolation would suggest that over 1,000 new scales join industrial/organizational psychology and organizational behavior in such a span. As scholars seek to incorporate those scales into their own research agendas, the overriding question becomes whether the scales allow for valid inferences to be made when using them. That question depends, in part, on the degree to which scale items adequately sample the universe of content associated with the construct (Cronbach, 1990; Nunnally, 1978).

Unfortunately, there are reasons to be concerned about the content of newly introduced scales, simply because so few go through any sort of content validation. As Table 1 reveals, content

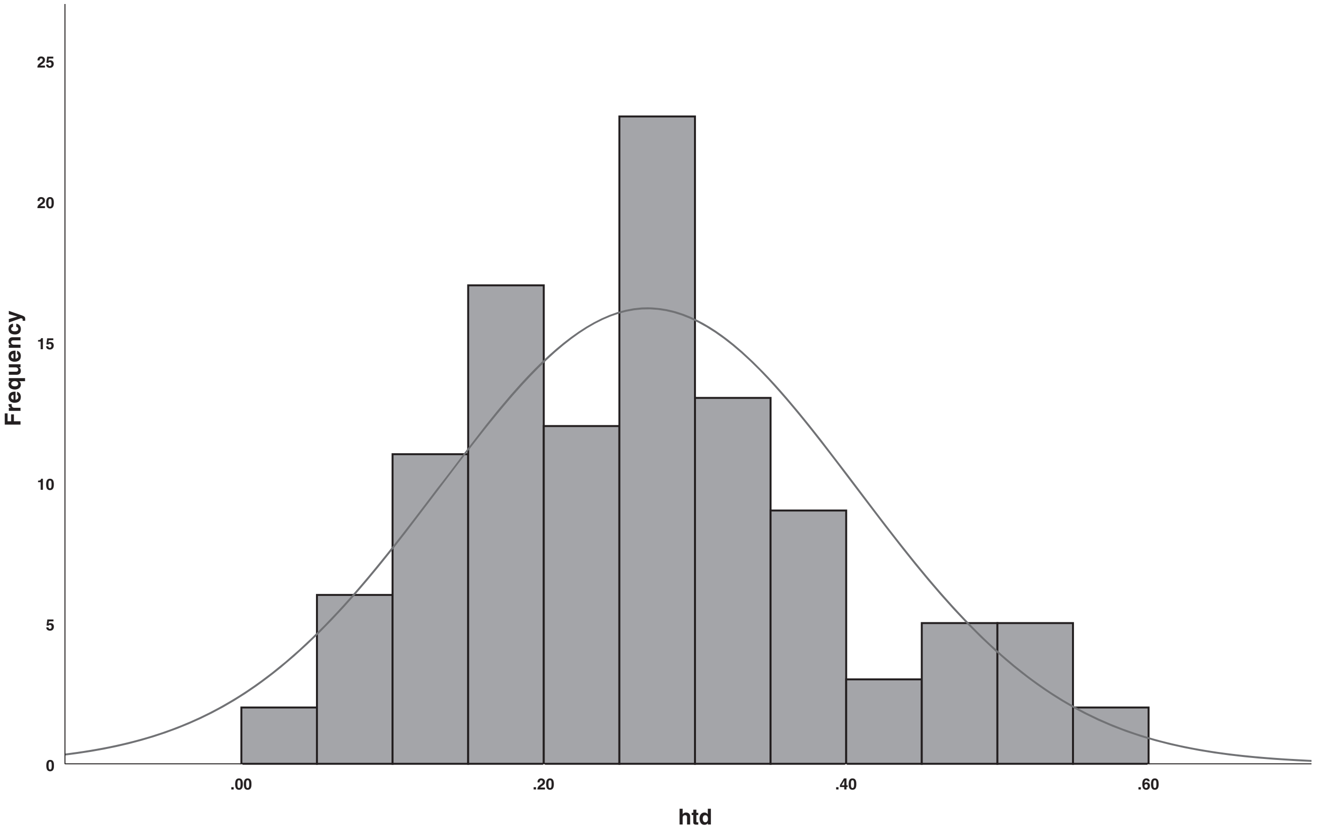


Figure 4. Distribution of *htd* for the 112 scales used to generate our evaluation criteria.

validation is the subject of much less discussion in *Journal of Applied Psychology* than, say, reliability and factor structure. That disparity can be seen in the articles we reviewed when generating our evaluation criteria. Only one of the 56 articles omitted information on reliability and only seven omitted information on factor structure. In contrast, 33 omitted discussion of content validation. Of course, the absence of such information does not mean that the 61 scales introduced in those 33 articles have content problems. It does, however, mean that reviewers may have been deprived of some key methodological details, as are potential users of those scales.

All that said, we are sensitive to the plight of authors who do need to create new scales for their work. Do they use one of the specific approaches spotlighted here for content validation, or do they merely borrow specific elements of those approaches? What are authors to make of the fact that Anderson and Gerbing's (1991) approach was only applied three times in our time window, with only one article using both of their recommended statistics (Spence et al., 2014)? Other questions surround Hinkin and Tracey's (1999) approach. Different authors have used varying numbers of anchors and varying wording of anchors, with few authors considering both definitional correspondence and definitional distinctiveness. Regardless of the choice of approach, authors also need to contend with questions about the number and nature of judges and how to actually execute the approaches within platforms like Qualtrics. Then, if all of those issues are navigated, one final question remains: How "good" is a c_{sv} of .67 anyway?

Our hope is that our evaluation criteria "demystify" many of those questions while providing standards for judging the results that flow from Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches. Appendix A and Appendix B illustrate exactly how to set up the approaches, including how to describe them to participants, how to display construct labels and definitions, what to say about reverse-worded items, how to provide a quiz to test understanding, and how to set up sorting boxes and rating scales. As the practical construction of these approaches becomes more accessible, their application within journals should increase.

Of course, a stronger trigger for such an increase may be the evaluation criteria in Table 5. Those guidelines possess a number of key strengths. For one, those guidelines follow LeBreton and

Table 3
Descriptives for Content Validation Statistics

Statistic	Range	Mean	Median	*SD*
p_{sa}	.24 to .98	.79	.82	.15
c_{sv}	−.52 to .95	.57	.63	.29
htc	.60 to .96	.87	.88	.06
htd	−.04 to .64	.27	.26	.14

Note. Descriptives for p_{sa}, c_{sv}, *htc*, and *htd* are based on a sample of 112 statistics.

Table 4
Correlations Between Scale Properties and Content Validation Statistics

Scale attributes	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. Number of items	4.79	1.67											
2. If reversed items used	.04	.21	.05										
3. Percent of reversed items	.03	.16	.04	.91*									
4. Coefficient alpha	.91	.05	.25*	−.06	.03								
5. If content validation reported	.46	.50	.03	−.20*	−.18	−.01							
6. If Anderson and Gerbing (1991) reported	.04	.21	.03	−.05	−.04	−.19*	.23*						
7. If Hinkin and Tracey (1999) reported	.18	.39	−.07	−.10	−.09	.21*	.50*	−.10					
8. Average *r* with orbiting scales	.43	.19	.14	−.02	−.05	−.08	.07	−.02	−.17				
9. p_{sa}	.79	.15	−.18	−.23*	−.12	.33*	−.06	−.16	.23*	−.55*			
10. c_{sv}	.57	.30	−.18	−.23*	−.13	.32*	−.06	−.16	.23*	−.55*	1.00*		
11. *htc*	.87	.06	−.26*	−.64*	−.64*	.24*	.08	−.14	.22*	−.18	.58*	.57*	
12. *htd*	.27	.14	−.07	−.19*	−.14	.20*	−.07	−.08	.17	−.66*	.74*	.75*	.40*

Note. $n = 112$ scales. Coefficient alpha values were taken from our correlation-based norming data.
* $p < .05$.

Senter (2008) in not being artificially dichotomous. Scale items do not possess (or not possess) definitional correspondence and definitional distinctiveness—both are a matter of degree. In addition, those guidelines are not arbitrary. They are rooted in results that seven years' worth of scales in four prominent journals earned, using a consistent set of instructions. Those guidelines are also derived from a pool of judges accessible to virtually all scholars—even scholars who lack research active faculty and who work at colleges and universities with no doctoral program, no subject pool, and small class sizes.

The guidelines in the top panel of Table 5 would allow a scholar to classify a c_{sv} of .67 as strong, insofar as it would fall in the 60th to 79th percentile of the 112 c_{sv} values earned by the scales in our data. The guidelines in the remaining panels of Table 5 would then

Table 5
Evaluation Criteria for Interpreting Content Validation Statistics

Percentile	Interpretation	p_{sa}	c_{sv}	*htc*	*htd*
Overall Criteria Not Normed to Average Correlation between Focal Scale and Orbiting Scales					
80th–99th	Very Strong	.91 and above	.81 and above	.91 and above	.35 and above
60th–79th	Strong	.82 to .91	.61 to .80	.87 to .90	.27 to .34
40th–59th	Moderate	.72 to .81	.51 to .60	.84 to .86	.18 to .26
20th–39th	Weak	.39 to .71	.05 to .50	.60 to .83	.04 to .17
0th–19th	Lack of	.38 and below	.04 and below	.59 and below	.03 and below
Stronger Average Correlation between Focal Scale and Orbiting Scales (.51 or above)					
80th–99th	Very Strong	.80 and above	.61 and above	.90 and above	.23 and above
60th–79th	Strong	.75 to .79	.50 to .60	.86 to .89	.15 to .22
40th–59th	Moderate	.60 to .74	.21 to .49	.82 to .85	.11 to .14
20th–39th	Weak	.24 to .59	.01 to .20	.63 to .81	.01 to .10
0th–19th	Lack of	.23 and below	.00 and below	.62 and below	.00 and below
More Moderate Average Correlation between Focal Scale and Orbiting Scales (.35 to .50)					
80th–99th	Very Strong	.91 and above	.83 and above	.92 and above	.34 and above
60th–79th	Strong	.81 to .90	.61 to .82	.89 to .91	.27 to .33
40th–59th	Moderate	.76 to .80	.52 to .60	.85 to .88	.20 to .26
20th–39th	Weak	.46 to .75	.01 to .51	.60 to .84	.09 to .19
0th–19th	Lack of	.45 and below	.00 and below	.59 and below	.08 and below
Weaker Average Correlation between Focal Scale and Orbiting Scales (.34 or below)					
80th–99th	Very Strong	.94 and above	.89 and above	.91 and above	.48 and above
60th–79th	Strong	.90 to .93	.80 to .87	.88 to .90	.35 to .47
40th–59th	Moderate	.84 to .89	.67 to .79	.86 to .87	.26 to .34
20th–39th	Weak	.52 to .83	.04 to .66	.67 to .85	.12 to .25
0th–19th	Lack of	.51 and below	.03 and below	.66 and below	.11 and below

Note. Percentiles for p_{sa}, c_{sv}, *htc*, and *htd* in the top panel of rows are based on quintiles of 112 statistics. The ranges for the average correlations of the remaining panels of rows between the focal scale and the orbiting scales are based on tertiles of those average correlations. The weaker range is the 0th–32nd percentile on those average correlations. The more moderate range is the 33rd–66th percentile on those average correlations. The stronger range is the 67th–99th percentile on those average correlations. Percentiles for p_{sa}, c_{sv}, *htc*, and *htd* are based on quintiles of 37, 38, and 37 statistics for the stronger, more moderate, and weaker ranges, respectively.

add vital nuance: if the correlations between the scale and its orbiting scales are stronger, then a c_{sv} of .67 would be classified as very strong. Taking into account the 37 c_{sv} values earned by the scales with stronger correlation values, that .67 would fall in the 80th to 99th percentile range. Alternatively, if the correlations between the scale and its orbiting scales are weaker, then a c_{sv} of .67 would instead be classified as only moderate. Taking into account the 37 c_{sv} values earned by the scales with weaker correlation values, that .67 would fall into the 40th to 59th percentile range. Such nuance is critical, as some content validation efforts will proceed in areas with higher intercorrelations (e.g., a new subfacet of an existing construct) while others will proceed in areas with lower intercorrelations (e.g., low base rate job behaviors).

The correlation-based criteria in Table 5 also reinforce how important the choice of orbiting scales is when applying Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches. Choosing orbiting scales that are strongly correlated with the focal scale results in very different guidelines than choosing orbiting scales that are weakly correlated with the focal scale. We would therefore reiterate the guidelines discussed earlier. Orbiting constructs should be venerable concepts in the literature with well understood definitions and commonly utilized scales. This guideline ensures that readers and reviewers have a strong understanding of the referents against which new scales are judged. Orbiting constructs should also represent "correlates" of the focal construct—existing at the same stage of causal flow—without having a "part-whole" relationship with it. Those guidelines balance the expected correlations between the focal construct and the orbiting ones, along with their expected distinctiveness. Along the same lines, choices should have the same referent as the focal construct. That guideline ensures that correlations and distinctiveness are driven by the substantive content of the scales rather than the entity being referenced in the items.

The focus on orbiting scales with Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches could provide an additional tool for combating construct proliferation. Shaffer et al. (2016) discuss the construct proliferation problem in the context of discriminant validation, reviewing a number of tools for performing such validation. Those tools include multitrait-multimethod matrices, constraining factor correlations to 1.0 in confirmatory factor analyses, and examining the size of disattenuated correlations. One challenge with such analyses, as noted by Shaffer et al. (2016), is that they depend on the quality of the items used to represent the construct. Two constructs may be identical but possess items that make them look distinct; two other constructs may be distinct but possess items that make them look identical. Although Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches were not referenced by Shaffer et al. (2016), both could help address that challenge. By incorporating *construct definitions* in a way that other tools do not, Anderson and Gerbing (1991) and Hinkin and Tracey (1999) add a definitional distinctiveness hurdle that lays apart from hurdles related to correlations among latent variables.

The sheer length of Table 2 illustrates the importance of having an additional tool in place to combat construct proliferation. If our extrapolation of 1,000 new scales joining industrial/organizational psychology and organizational behavior in a 7-year span is accurate, it seems certain that many of the constructs associated with such scales are not truly "new." Focusing specifically on Table 2, some new scales are merely applying existing constructs to new referents (e.g., career satisfaction, coworker emotional support, customer unethical behavior, family incivility) or splitting existing constructs into more specific facets (e.g., constructive voice, interest/enjoyment scales self-efficacy, personal-to-work conflict). Other new scales do represent ostensibly new constructs, however, from bottom-line mentality to naiveté to wanderlust. As future reviewers examine additional constructs that are introduced to the literature, they can both expect the application of Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches and guide the application of them during revisions. For example, reviewers can suggest additional orbiting scales that might not have been originally included, or require more stringent interpretations of the evaluation criteria in Table 5.

Finally, as new scales are evaluated against our guidelines, it is worth considering the issue of reverse-worded items. Theoretically, either Anderson and Gerbing's (1991) approach or Hinkin and Tracey's (1999) approach should work with such items, and the former used such items when showcasing the method. Still, it is easy to see why a participant might view "I work hard in my job" as more corresponding to "Motivation: The effort expended in relation to work" than "I often feel lazy at the office." The positive pole of a construct's continuum will often seem more akin to how a construct definition is worded. To some extent, our warnings about the potential impact of reverse-worded items on p_{sa}, c_{sv}, *htc*, and *htd* echo warnings about their impact on factor analytic results (Schmitt & Stults, 1985; Schriesheim & Eisenbach, 1995). In cases where reverse-worded items are deemed critical to a scale, we see two potential options to consider. First, the instructions (and associated quiz) could be made even more explicit about this issue. Second, the construct definition could itself be rewritten in a way that alludes to both the positive and negative pole of the construct. For example, the definition above could be rewritten as "Motivation: The effort expended (or not expended) in relation to work."

Limitations

Our work includes a number of limitations that should be noted. First and foremost, our evaluation criteria are bounded in Likert-style scales of explicit constructs. Our work cannot speak to the content validation of scales that possess a forced choice or paired comparison format. For example, Meglino, Ravlin, and Adkins (1989) measured values about achievement, concern, fairness, and honesty with a forced choice format that asked participants to put more emphasis on one value over another within each item. Our work also cannot speak to the content validation of implicit measures and tests that possess a particular scoring key. For example, James and McIntyre (2000) measured aggression with a conditional reasoning test where participants thought their way through certain questions, with some options representing justifications for aggression. As another example, Lievens and Sackett (2012) assessed interpersonal skills with a situational judgment test where participants reacted to critical incidents with multiple choice questions about effective responses. We suspect these sorts of measures represent cases where expert raters are more appropriate than naïve raters. After all, judging the content of such measures requires more than intellectual ability or linguistic skills. Some particular awareness of the content domain is necessary, as is some familiarity with the technical aspects of the measurement technique.

An additional limitation also concerns the use of expert judges versus naïve judges. Our focus was solely on whether scales adequately sample the universe of content associated with a construct. Some contexts bring with them additional criteria for judging measurement. For example, Lawshe's (1975) approach to content validation was rooted in concerns about the legality of selection measures. He noted that those in charge of such efforts often have to serve as expert witnesses in discrimination cases, either as defendant or plaintiff. It may be that expert judges will seem more compelling than naïve judges in such contexts, especially to a jury. If so, that context would argue for an approach that is rooted more in Lawshe's (1975) mechanics than either Anderson and Gerbing's (1991) or Hinkin and Tracey's (1999).

Even outside of legal settings, some scholars may find expert judges more compelling as a matter of course than naïve judges, even given Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) views on the subject. Indeed, we were struck by how many of the articles that conducted some kind of validation used doctoral students as judges (Erdogan et al., 2015; Harrison & Wagner, 2016; Leroy et al., 2015; Liang et al., 2012; Zhang et al., 2015), or a combination of doctoral students and faculty (Fast et al., 2014; Klein et al., 2014; Shepherd et al., 2011; Wilson & Baumann, 2015). The appeal of such judges is straightforward, given that they bring 2 years of coursework, the experience of studying for comprehensive exams, and several years of experience reading articles, writing articles, and attending conferences into their sorting and rating tasks. Doctoral students and faculty can therefore look beyond the idiosyncratic structure of a definition—or an item phrasing with multiple potential meanings—by drawing on their experience and expertise.

Given such differences between expert raters and naïve raters, we caution that the criteria in Table 5 should not be applied to applications of Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches that use doctoral students or faculty. Both approaches were geared around a literal reading of a construct definition—not a reading filtered through one's own subject matter expertise. Indeed, one could envision cases where doctoral students or faculty could react to orbiting scale items with an immediate awareness of the scale's label and creators. For example, a doctoral participant in our study who was rating items for constructive voice (Maynes & Podsakoff, 2014), in-role behavior (Williams & Anderson, 1991), and civic virtue (Podsakoff, MacKenzie, Moorman, & Fetter, 1990) against the constructive voice definition might instantly recognize "I adequately complete assigned duties" as a Williams and Anderson (1991) item, given previous use of that scale in his or her research. That participant might not actually engage in the linguistic comparison between "adequately completing assigned duties" and "voluntarily expressing ideas, information, or opinions focused on effecting organizationally functional change to the work context" because the item has been recognized as a distractor. That difference in the rating and sorting experience could result in different results—especially for item distinctiveness.

Finally, when considering the degree to which scale items adequately sample the universe of content associated with a construct, there is one aspect that is not tapped by the statistics benchmarked here. As noted in the opening of our paper, that aspect is deficiency—the degree to which items fail to capture some part of the content universe (Cronbach, 1990; Nunnally, 1978). Neither Anderson and Gerbing's (1991) approach nor Hinkin and Tracey's (1999) approach speak to this critical consideration in content validation efforts. It is therefore important to emphasize that adequate p_{sa}, c_{sv}, *htc*, and *htd* values are not—in and of themselves—sufficient metrics for judging the quality of a scale. Much like coefficient alpha, average factor loadings, or the fit of a measurement model, they tell only one piece of a larger content validation story.

Coming up with a quantitative means of judging deficiency therefore stands as a valuable direction for future research. Theoretically, raters could be presented with a construct definition and a set of items and be asked to judge the degree to which the items capture the full universe of content included within that definition. Some variation of the scale in Appendix B could be used in this regard, with anchors such as 1 = *items do an extremely bad job of capturing the full universe of content associated with the definition* to 7 = *items do an extremely good job of capturing the full universe of content associated with the definition*. Raters could also be asked to suggest additional items that seem distinct from the items provided while still matching the construct definition. Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approach could even be used as a follow-up to gauge the definitional correspondence of those additional items. Returning to our prior discussion, this is almost certainly a case where expert raters would be preferable to naïve raters—given the need for extensive knowledge of the content domain.

Although we see a great deal of promise in this sort of approach, it will likely require more detailed and faceted construct definitions than are often provided by authors. For example, leader–member exchange quality can be defined as the degree to which the leader–follower relationship is developmentally mature and effective (Graen & Uhl-Bien, 1995). It can also be defined as the degree to which the leader-follower relationship is characterized by mutual respect, trust, and obligation (Graen & Uhl-Bien, 1995). The rating and suggestion task described above would be much more challenging with the more global leader-member exchange definition than with the more faceted one, given that the facets bring both detail and boundary to the content universe. We therefore suggest that authors of future scales be sure to provide detailed and faceted construct definitions in place of—or alongside—more global definitions. Those definitions could then be given to expert raters to perform the kind of assessment described above. That approach would take advantage of their deeper content expertise and broader training, relative to naïve raters.

Conclusion

Constructs obviously lay at the core of theorizing in industrial/organizational psychology and organizational behavior (Bacharach, 1989; Whetten, 1989). Writing items that tap those constructs, in turn, lays at the core of measurement (Hendrick et al., 2013; Hinkin, 1995, 1998; MacKenzie et al., 2011). There is something elegantly simple about Anderson and Gerbing's (1991) and Hinkin and Tracey's (1999) approaches to gauging whether items correspond to a construct's definition—and do so more strongly than to other definitions. Given that simplicity, it seems surprising that the approaches are used so infrequently in the four journals we examined. We hope that providing these evaluation criteria—and providing detailed, transparent, and replicable instructions for the approaches—will increase the extent to which they are used to support new scale measures.

References

References marked with an asterisk indicate studies included in the generation of our evaluation criteria.

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies, 51,* 143–174. http://dx.doi.org/10.1111/joms.12058

Aguinis, H., & Vandenberg, R. J. (2014). An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior, 1,* 569–595. http://dx.doi.org/10.1146/annurev-orgpsych-031413-091231

Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology, 76,* 732–740. http://dx.doi.org/10.1037/0021-9010.76.5.732

Arnold, J. A., Arad, S., Rhoades, J. A., & Drasgow, F. (2000). The empowering leadership questionnaire: The construction and validation of a new scale for measuring leader behaviors. *Journal of Organizational Behavior, 21,* 249–269. http://dx.doi.org/10.1002/(SICI)1099-1379(200005)21:3<249::AID-JOB10>3.0.CO;2-#

Ashford, S. J. (1986). Feedback-seeking in individual adaptation: A resource perspective. *Academy of Management Journal, 29,* 465–487.

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review, 14,* 496–515. http://dx.doi.org/10.5465/amr.1989.4308374

Bacharach, S. B., Bamberger, P. A., & Conley, S. C. (1990). Work processes, role conflict, and role overload: The case of nurses and engineers in the public sector. *Work and Occupations, 17,* 199–228. http://dx.doi.org/10.1177/0730888490017002004

*Baer, M. D., Dhensa-Kahlon, R. K., Colquitt, J. A., Rodell, J. B., Outlaw, R., & Long, D. M. (2015). Uneasy lies the head that bears the trust: The effects of feeling trusted on emotional exhaustion. *Academy of Management Journal, 58,* 1637–1657. http://dx.doi.org/10.5465/amj.2014.0246

*Barasch, A., Levine, E. E., & Schweitzer, M. E. (2016). Bliss is ignorance: How the magnitude of expressed happiness influences perceived naiveté and interpersonal exploitation. *Organizational Behavior and Human Decision Processes, 137,* 184–206. http://dx.doi.org/10.1016/j.obhdp.2016.05.006

Bass, B. M., & Avolio, B. J. (1990). *Multifactor leadership questionnaire*. Palo Alto, CA: Consulting Psychologists Press.

*Belmi, P., & Neale, M. (2014). Mirror, mirror on the wall, who's the fairest of them all? Thinking that one is attractive increases the tendency to support inequality. *Organizational Behavior and Human Decision Processes, 124,* 133–149. http://dx.doi.org/10.1016/j.obhdp.2014.03.002

*Belmi, P., & Pfeffer, J. (2016). Power and death: Mortality salience increases power seeking while feeling powerful reduces death anxiety. *Journal of Applied Psychology, 101,* 702–720. http://dx.doi.org/10.1037/apl0000076

*Bennett, A. A., Gabriel, A. S., Calderwood, C., Dahling, J. J., & Trougakos, J. P. (2016). Better together? Examining profiles of employee recovery experiences. *Journal of Applied Psychology, 101,* 1635–1654. http://dx.doi.org/10.1037/apl0000157

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology, 85,* 349–360. http://dx.doi.org/10.1037/0021-9010.85.3.349

*Bindl, U. K., Parker, S. K., Totterdell, P., & Hagger-Johnson, G. (2012). Fuel of the self-starter: How mood relates to proactive goal regulation. *Journal of Applied Psychology, 97,* 134–150. http://dx.doi.org/10.1037/a0024368

*Bolino, M. C., Hsiung, H. H., Harvey, J., & LePine, J. A. (2015). "Well, I'm tired of tryin'!" Organizational citizenship behavior and citizenship fatigue. *Journal of Applied Psychology, 100,* 56–74. http://dx.doi.org/10.1037/a0037583

Brockner, J., Siegel, P. A., Daly, J. P., Tyler, T., & Martin, C. (1997). When trust matters: The moderating effect of outcome favorability. *Administrative Science Quarterly, 42,* 558–583. http://dx.doi.org/10.2307/2393738

*Burris, E. R. (2012). The risks and rewards of speaking up: Managerial responses to employee voice. *Academy of Management Journal, 55,* 851–875. http://dx.doi.org/10.5465/amj.2010.0562

*Cable, D. M., & Kay, V. S. (2012). Striving for self-verification during organizational entry. *Academy of Management Journal, 55,* 360–380. http://dx.doi.org/10.5465/amj.2010.0397

Cammann, C., Fichman, M., Jenkins, G. D., Jr., & Klesh, J. R. (1983). Assessing the attitudes and perceptions of organizational members. In S. E. Seashore, E. E. Lawler, P. H. Mirvis, & C. Cammann (Eds.), *Assessing organizational change* (pp. 71–138). New York, NY: Wiley and Sons.

Carr, J. C., Boyar, S. L., & Gregory, B. T. (2008). The moderating effect of work-family centrality on work-family conflict, organizational attitudes, and turnover behavior. *Journal of Management, 34,* 244–262. http://dx.doi.org/10.1177/0149206307309262

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology, 56,* 267–283. http://dx.doi.org/10.1037/0022-3514.56.2.267

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods, 4,* 62–83. http://dx.doi.org/10.1177/109442810141004

Chiu, C. Y., Hong, Y. Y., & Dweck, C. S. (1997). Lay dispositionism and implicit theories of personality. *Journal of Personality and Social Psychology, 73,* 19–30. http://dx.doi.org/10.1037/0022-3514.73.1.19

*Colbert, A. E., Bono, J. E., & Purvanova, R. K. (2016). Flourishing via workplace relationships: Moving beyond instrumental support. *Academy of Management Journal, 59,* 1199–1223. http://dx.doi.org/10.5465/amj.2014.0506

*Colquitt, J. A., Baer, M. D., Long, D. M., & Halvorsen-Ganepola, M. D. K. (2014). Scale indicators of social exchange relationships: A comparison of relative content validity. *Journal of Applied Psychology, 99,* 599–618. http://dx.doi.org/10.1037/a0036374

*Colquitt, J. A., Long, D. M., Rodell, J. B., & Halvorsen-Ganepola, M. D. K. (2015). Adding the "in" to justice: A qualitative and quantitative investigation of the differential effects of justice rule adherence and violation. *Journal of Applied Psychology, 100,* 278–294. http://dx.doi.org/10.1037/a0038131

Conger, J. A., & Kanungo, R. N. (1994). Charismatic leadership in organizations: Perceived behavioral attributes and their measurement. *Journal of Organizational Behavior, 15,* 439–452. http://dx.doi.org/10.1002/job.4030150508

Cronbach, L. J. (1990). *Essentials of psychological testing*. New York, NY: Harper & Row.

Dang, C. T., Umphress, E. E., & Mitchell, M. S. (2017). Leader social accounts of subordinates' unethical behavior: Examining observer reactions to leader social accounts with moral disengagement language. *Journal of Applied Psychology, 102,* 1448–1461. http://dx.doi.org/10.1037/apl0000233

Djurdjevic, E., Stoverink, A. C., Klotz, A. C., Koopman, J., da Motta Veiga, S. P., Yam, K. C., & Chiang, J. T.-J. (2017). Workplace status: The development and validation of a scale. *Journal of Applied Psychology, 102,* 1124–1147. http://dx.doi.org/10.1037/apl0000202

*Dobrow, S. R., & Tosti-Kharas, J. (2011). Calling: The development of a scale measure. *Personnel Psychology, 64,* 1001–1049. http://dx.doi.org/10.1111/j.1744-6570.2011.01234.x

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18,* 192–203. http://dx.doi.org/10.1037/1040-3590.18.2.192

Eisenberger, R., Armeli, S., Rexwinkel, B., Lynch, P. D., & Rhoades, L. (2001). Reciprocation of perceived organizational support. *Journal of Applied Psychology, 86,* 42–51. http://dx.doi.org/10.1037/0021-9010.86.1.42

*Eisenberger, R., Karagonlar, G., Stinglhamber, F., Neves, P., Becker, T. E., Gonzalez-Morales, M. G., & Steiger-Mueller, M. (2010). Leader-member exchange and affective organizational commitment: The contribution of supervisor's organizational embodiment. *Journal of Applied Psychology, 95,* 1085–1103. http://dx.doi.org/10.1037/a0020858

*Erdogan, B., Bauer, T. N., & Walter, J. (2015). Deeds that help and words that hurt: Helping and gossip as moderators of the relationship between leader–member exchange and advice network centrality. *Personnel Psychology, 68,* 185–214. http://dx.doi.org/10.1111/peps.12075

*Fast, N. J., Burris, E. R., & Bartel, C. A. (2014). Managing to stay in the dark: Managerial self-efficacy, ego defensiveness, and the aversion to employee voice. *Academy of Management Journal, 57,* 1013–1034. http://dx.doi.org/10.5465/amj.2012.0393

*Gelfand, M. J., Leslie, L. M., Keller, K., & de Dreu, C. (2012). Conflict cultures in organizations: How leaders shape conflict cultures and their organizational-level consequences. *Journal of Applied Psychology, 97,* 1131–1147. http://dx.doi.org/10.1037/a0029993

*Gonzalez-Gomez, H. V., & Richter, A. W. (2015). Turning shame into creativity: The importance of exposure to creative team environments. *Organizational Behavior and Human Decision Processes, 126,* 142–161. http://dx.doi.org/10.1016/j.obhdp.2014.09.004

Graen, G. B., & Uhl-Bien, M. (1995). Relationship-based approach to leadership: Development of leader-member exchange (LMX) theory of leadership over 25 years: Applying a multi-level multi-domain perspective. *The Leadership Quarterly, 6,* 219–247. http://dx.doi.org/10.1016/1048-9843(95)90036-5

Grant, A. M., Dutton, J. E., & Rosso, B. D. (2008). Giving commitment: Employee support programs and the prosocial sensemaking process. *Academy of Management Journal, 51,* 898–918. http://dx.doi.org/10.5465/amj.2008.34789652

*Greenbaum, R. L., Mawritz, M. B., & Eissa, G. (2012). Bottom-line mentality as an antecedent of social undermining and the moderating roles of core self-evaluations and conscientiousness. *Journal of Applied Psychology, 97,* 343–359. http://dx.doi.org/10.1037/a0025217

*Greenbaum, R. L., Quade, M. J., Mawritz, M. B., Kim, J., & Crosby, D. (2014). When the customer is unethical: The explanatory role of employee emotional exhaustion onto work-family conflict, relationship conflict with coworkers, and job neglect. *Journal of Applied Psychology, 99,* 1188–1203. http://dx.doi.org/10.1037/a0037221

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50,* 327–347. http://dx.doi.org/10.5465/amj.2007.24634438

*Guillaume, Y. R. F., Knippenberg, D. V., & Brodbeck, F. C. (2014). Nothing succeeds like moderation: A social self-regulation perspective on cultural dissimilarity and performance. *Academy of Management Journal, 57,* 1284–1308. http://dx.doi.org/10.5465/amj.2013.0046

*Gupta, N., Ganster, D. C., & Kepes, S. (2013). Assessing the validity of sales self-efficacy: A cautionary tale. *Journal of Applied Psychology, 98,* 690–700. http://dx.doi.org/10.1037/a0032232

Gutek, B. A., Searle, S., & Klepa, L. (1991). Rational versus gender role explanations for work-family conflict. *Journal of Applied Psychology, 76,* 560–568. http://dx.doi.org/10.1037/0021-9010.76.4.560

*Hannah, S. T., Jennings, P. L., Bluhm, D., Peng, A. C., & Schaubroeck, J. M. (2014). Duty orientation: Theoretical development and preliminary construct testing. *Organizational Behavior and Human Decision Processes, 123,* 220–238. http://dx.doi.org/10.1016/j.obhdp.2013.10.007

*Harrison, S. H., & Wagner, D. T. (2016). Spilling outside the box: The effects of individuals' creative behaviors at work on time spent with their spouses at home. *Academy of Management Journal, 59,* 841–859. http://dx.doi.org/10.5465/amj.2013.0560

Hendrick, T. A. M., Fischer, A. R. H., Tobi, H., & Frewer, L. J. (2013). Self-reported attitude scales: Current practice in adequate assessment of reliability, validity, and dimensionality. *Journal of Applied Social Psychology, 43,* 1538–1552. http://dx.doi.org/10.1111/jasp.12147

Hideg, I., & Ferris, D. L. (2017). Dialectical thinking and fairness-based perspectives of affirmative action. *Journal of Applied Psychology, 102,* 782–801. http://dx.doi.org/10.1037/apl0000207

Hinkin, T. R. (1985). *Development and application of new social power measures in superior-subordinate relationships* (Unpublished doctoral dissertation). University of Florida, Gainesville, FL.

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

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

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods, 2,* 175–186. http://dx.doi.org/10.1177/109442819922004

Indvik, J. (1985). *A path-goal theory investigation of superior subordinate relationships* (Unpublished doctoral dissertation). University of Wisconsin, Madison, WI.

James, L. R., & McIntyre, M. D. (2000). *Conditional Reasoning Test of Aggression test manual*. Knoxville, TN: Innovative Assessment Technology.

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment, 22,* 420–432. http://dx.doi.org/10.1037/a0019265

*Klein, H. J., Cooper, J. T., Molloy, J. C., & Swanson, J. A. (2014). The assessment of commitment: Advantages of a unidimensional, target-free approach. *Journal of Applied Psychology, 99,* 222–238. http://dx.doi.org/10.1037/a0034751

*Kraimer, M. L., Seibert, S. E., Wayne, S. J., Liden, R. C., & Bravo, J. (2011). Antecedents and outcomes of organizational support for development: The critical role of career opportunities. *Journal of Applied Psychology, 96,* 485–500. http://dx.doi.org/10.1037/a0021452

*Kraimer, M. L., Shaffer, M. A., Harrison, D. A., & Ren, H. (2012). No place like home? An identity strain perspective on repatriate turnover. *Academy of Management Journal, 55,* 399–420. http://dx.doi.org/10.5465/amj.2009.0644

Landis, B., Kilduff, M., Menges, J. I., & Kilduff, G. J. (2018). The paradox of agency: Feeling powerful reduces brokerage opportunity recognition yet increases willingness to broker. *Journal of Applied Psychology, 103,* 929–938. http://dx.doi.org/10.1037/apl0000299

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28,* 563–575. http://dx.doi.org/10.1111/j.1744-6570.1975.tb01393.x

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852. http://dx.doi.org/10.1177/1094428106296642

Lehman, W. E. K., & Simpson, D. D. (1992). Employee substance use and on-the-job behaviors. *Journal of Applied Psychology, 77,* 309–321. http://dx.doi.org/10.1037/0021-9010.77.3.309

*Leroy, S., Shipp, A. J., Blount, A., & Licht, J. G. (2015). Synchrony preference: Why some people go with the flow and some don't. *Personnel Psychology, 68,* 759–809. http://dx.doi.org/10.1111/peps.12093

*Leslie, L. M., Manchester, C. F., Park, T., & Mehng, S. A. (2012). Flexible work practices: A source of career premiums or penalties? *Academy of Management Journal, 55,* 1407–1428. http://dx.doi.org/10.5465/amj.2010.0651

Levenson, H. (1981). Differentiating among internality, powerful others, and chance. In H. M. Lefcourt (Ed.), *Research with the locus of control construct* (Vol. 1, pp. 15–63). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-443201-7.50006-3

*Liang, J., Farh, C. I. C., & Farh, J. L. (2012). Psychological antecedents of promotive and prohibitive voice: A two-wave examination. *Academy of Management Journal, 55,* 71–92. http://dx.doi.org/10.5465/amj.2010.0176

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology, 97,* 460–468. http://dx.doi.org/10.1037/a0025741

*Lim, S., & Tai, K. (2014). Family incivility and job performance: A moderated mediation model of psychological distress and core self-evaluation. *Journal of Applied Psychology, 99,* 351–359. http://dx.doi.org/10.1037/a0034486

*Lin, S.-H., Ma, J., & Johnson, R. E. (2016). When ethical leader behavior breaks bad: How ethical leader behavior can turn abusive via ego depletion and moral licensing. *Journal of Applied Psychology, 101,* 815–830. http://dx.doi.org/10.1037/apl0000098

*Long, C. P., Bendersky, C., & Morrill, C. (2011). Fairness monitoring: Linking managerial controls and fairness judgments in organizations. *Academy of Management Journal, 54,* 1045–1068. http://dx.doi.org/10.5465/amj.2011.0008

*Long, D. M., Baer, M. D., Colquitt, J. A., Outlaw, R., & Dhensa-Kahlon, R. K. (2015). What will the boss think? The impression management implications of supportive relationships with star and project peers. *Personnel Psychology, 68,* 463–498. http://dx.doi.org/10.1111/peps.12091

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *Management Information Systems Quarterly, 35,* 293–334. http://dx.doi.org/10.2307/23044045

Mael, F., & Ashforth, B. E. (1992). Alumni and their alma mater: A partial test of the reformulated model of organizational identification. *Journal of Organizational Behavior, 13,* 103–123. http://dx.doi.org/10.1002/job.4030130202

*May, D. R., Chang, Y. K., & Shao, R. (2015). Does ethical membership matter? Moral identification and its organizational implications. *Journal of Applied Psychology, 100,* 681–694. http://dx.doi.org/10.1037/a0038344

*Mayer, D. M., Thau, S., Workman, K. M., Van Dijke, M., & De Cremer, D. (2012). Leader mistreatment, employee hostility, and deviant behaviors: Integrating self-uncertainty and thwarted needs perspectives on deviance. *Organizational Behavior and Human Decision Processes, 117,* 24–40. http://dx.doi.org/10.1016/j.obhdp.2011.07.003

*Maynes, T. D., & Podsakoff, P. M. (2014). Speaking more broadly: An examination of the nature, antecedents, and consequences of an expanded set of employee voice behaviors. *Journal of Applied Psychology, 99,* 87–112. http://dx.doi.org/10.1037/a0034284

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17,* 437–455. http://dx.doi.org/10.1037/a0028085

Meglino, B. M., Ravlin, E. C., & Adkins, C. L. (1989). A work values approach to corporate culture: A field test of the value congruence process and its relationship to individual outcomes. *Journal of Applied Psychology, 74,* 424–432. http://dx.doi.org/10.1037/0021-9010.74.3.424

*Mitchell, M. S., Vogel, R. M., & Folger, R. (2015). Third parties' reactions to the abusive supervision of coworkers. *Journal of Applied Psychology, 100,* 1040–1055. http://dx.doi.org/10.1037/apl0000002

*Mohammed, S., & Nadkarni, S. (2011). Temporal diversity and team performance: The moderating role of team temporal leadership. *Academy of Management Journal, 54,* 489–508. http://dx.doi.org/10.5465/amj.2011.61967991

*Moore, C., Detert, J. R., Trevino, L. K., Baker, V. L., & Mayer, D. M. (2012). Why employees do bad things: Moral disengagement and unethical organizational behavior. *Personnel Psychology, 65,* 1–48. http://dx.doi.org/10.1111/j.1744-6570.2011.01237.x

Moorman, R. H., & Blakely, G. L. (1995). Individualism-collectivism as an individual difference predictor of organizational citizenship behavior. *Journal of Organizational Behavior, 16,* 127–142. http://dx.doi.org/10.1002/job.4030160204

Motowidlo, S. J., Packard, J. S., & Manning, M. R. (1986). Occupational stress: Its causes and consequences for job performance. *Journal of Applied Psychology, 71,* 618–629. http://dx.doi.org/10.1037/0021-9010.71.4.618

Nielsen, I. K., Jex, S. M., & Adams, G. A. (2000). Development and validation of scores on a two-dimensional workplace friendship scale. *Educational and Psychological Measurement, 60,* 628–643. http://dx.doi.org/10.1177/00131640021970655

*Nifadkar, S., Tsui, A. S., & Ashforth, B. E. (2012). The way you make me feel and behave: Supervisor-triggered newcomer affect and approach-avoidance behavior. *Academy of Management Journal, 55,* 1146–1168. http://dx.doi.org/10.5465/amj.2010.0133

*Nishii, L. A. (2013). The benefits of climate for inclusion for gender-diverse groups. *Academy of Management Journal, 56,* 1754–1774. http://dx.doi.org/10.5465/amj.2009.0823

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Oldham, G. R., & Cummings, A. (1996). Employee creativity: Personal and contextual factors at work. *Academy of Management Journal, 39,* 607–634.

*Owens, B. P., Baker, W. E., Sumpter, D. M., & Cameron, K. S. (2016). Relational energy at work: Implications for job engagement and job performance. *Journal of Applied Psychology, 101,* 35–49. http://dx.doi.org/10.1037/apl0000032

Owens, B. P., Johnson, M. D., & Mitchell, T. R. (2013). Expressed humility in organizations: Implications for performance, teams, and leadership. *Organization Science, 24,* 1517–1538. http://dx.doi.org/10.1287/orsc.1120.0795

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23,* 184–188. http://dx.doi.org/10.1177/0963721414531598

Peterson, M. F., Smith, P. B., Akande, A., Ayestaran, S., Bochner, S., Callan, V., . . . Viedge, C. (1995). Role conflict, ambiguity, and overload: A 21-nation study. *Academy of Management Journal, 38,* 429–452.

*Plouffe, C. R., & Gregoire, Y. (2011). Intraorganizational employee navigation and socially derived outcomes: Conceptualization, validation, and effects on overall performance. *Personnel Psychology, 64,* 693–738. http://dx.doi.org/10.1111/j.1744-6570.2011.01223.x

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903. http://dx.doi.org/10.1037/0021-9010.88.5.879

Podsakoff, P. M., MacKenzie, S. B., Moorman, R. H., & Fetter, R. (1990). Transformational leader behaviors and their effects on followers' trust in leader, satisfaction, and organizational citizenship behaviors. *The Leadership Quarterly, 1,* 107–142. http://dx.doi.org/10.1016/1048-9843(90)90009-7

*Qin, X., Ren, R., Zhang, Z. X., & Johnson, R. E. (2015). Fairness heuristics and substitutability effects: Inferring the fairness of outcomes, procedures, and interpersonal treatment when employees lack clear information. *Journal of Applied Psychology, 100,* 749–766. http://dx.doi.org/10.1037/a0038084

*Rodell, J. B. (2013). Finding meaning through volunteering: Why do employees volunteer and what does it mean for their jobs? *Academy of Management Journal, 56,* 1274–1294. http://dx.doi.org/10.5465/amj.2012.0611

*Rodell, J. B., & Lynch, J. W. (2016). Perceptions of employee volunteering: Is it "credited" or "stigmatized" by colleagues? *Academy of Management Journal, 59,* 611–635. http://dx.doi.org/10.5465/amj.2013.0566

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press. http://dx.doi.org/10.1515/9781400876136

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9,* 367–373. http://dx.doi.org/10.1177/014662168500900405

Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management, 21,* 1177–1193. http://dx.doi.org/10.1177/014920639502100609

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,* 385–417. http://dx.doi.org/10.1177/014920639301900208

Schwartz, S. H. (1994). Are there universal aspects int he structure and contents of human values? *Journal of Social Issues, 50,* 19–45. http://dx.doi.org/10.1111/j.1540-4560.1994.tb01196.x

Seibert, S. E., Crant, J. M., & Kraimer, M. L. (1999). Proactive personality and career success. *Journal of Applied Psychology, 84,* 416–427. http://dx.doi.org/10.1037/0021-9010.84.3.416

*Seibert, S. E., Kraimer, M. L., Holtom, B. C., & Pierotti, A. J. (2013). Even the best laid plans sometimes go askew: Career self-management processes, career shocks, and the decision to pursue graduate education. *Journal of Applied Psychology, 98,* 169–182. http://dx.doi.org/10.1037/a0030882

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,* 80–110. http://dx.doi.org/10.1177/1094428115598239

Sheehan, K. B., & Pittman, M. (2016). *Amazon's Mechanical Turk for academics: The HIT handbook for social science research*. Irvine, CA: Melvin & Leigh.

*Shepherd, D. A., Patzelt, H., & Wolfe, M. (2011). Moving forward from project failure: Negative emotions, affective commitment, and learning from the experience. *Academy of Management Journal, 54,* 1229–1259. http://dx.doi.org/10.5465/amj.2010.0102

Spector, P. E. (1985). Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey. *American Journal of Community Psychology, 13,* 693–713. http://dx.doi.org/10.1007/BF00929796

*Spence, J. R., Brown, D. J., Keeping, L. M., & Lian, H. (2014). Helpful today, but not tomorrow? Feeling grateful as a predictor of daily organizational citizenship behaviors. *Personnel Psychology, 67,* 705–738.

Spreitzer, G. M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal, 38,* 1442–1465.

Stogdill, R. M. (1963). *Manual for the leader behavior description questionnaire-Form XII*. Columbus, OH: Bureau of Business Research, Ohio State University.

Stone-Romero, E. F. (1994). Construct validity issues in organizational behavior research. In J. Greenberg (Ed.), *Organizational behavior: The state of the science* (pp. 155–179). Hillsdale, NJ: Erlbaum.

*Sung, S. Y., & Choi, J. N. (2012). Effects of team knowledge management on the creativity and financial performance of organizational teams. *Organizational Behavior and Human Decision Processes, 118,* 4–13. http://dx.doi.org/10.1016/j.obhdp.2012.01.001

Susskind, A. M., Kacmar, K. M., & Borchgrevink, C. P. (2003). Customer service providers' attitudes relating to customer service and customer satisfaction in the customer-server exchange. *Journal of Applied Psychology, 88,* 179–187. http://dx.doi.org/10.1037/0021-9010.88.1.179

*Tepper, B. J., Mitchell, M. S., Haggard, D. L., Kwan, H. K., & Park, H. M. (2015). On the exchange of hostility with supervisors: An examination of self-enhancing and self-defeating perspectives. *Personnel Psychology, 68,* 723–758. http://dx.doi.org/10.1111/peps.12094

*Tinsley, C. H., Howell, T. M., & Amanatullah, E. T. (2015). Who should bring home the bacon? How deterministic views of gender constrain spousal wage preferences. *Organizational Behavior and Human Decision Processes, 126,* 37–48. http://dx.doi.org/10.1016/j.obhdp.2014.09.003

Ullrich, J., Christ, O., & van Dick, R. (2009). Substitutes for procedural fairness: Prototypical leaders are endorsed whether they are fair or not. *Journal of Applied Psychology, 94,* 235–244. http://dx.doi.org/10.1037/a0012936

Vandewalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement, 57,* 995–1015. http://dx.doi.org/10.1177/0013164497057006009

van Dijke, M., De Cremer, D., Langendijk, G., & Anderson, C. (2018). Ranking low, feeling high: How hierarchical position and experienced power promote prosocial behavior in response to procedural justice. *Journal of Applied Psychology, 103,* 164–181. http://dx.doi.org/10.1037/apl0000260

Van Dyne, L., & LePine, J. A. (1998). Helping and voice extra-role behaviors: Evidence of construct and predictive validity. *Academy of Management Journal, 41,* 108–119.

Van Katwyk, P. T., Fox, S., Spector, P. E., & Kelloway, E. K. (2000). Using the Job-Related Affective Well-Being Scale (JAWS) to investigate affective responses to work stressors. *Journal of Occupational Health Psychology, 5,* 219–230. http://dx.doi.org/10.1037/1076-8998.5.2.219

*Wang, X. H., & Howell, J. M. (2010). Exploring the dual-level effects of transformational leadership on followers. *Journal of Applied Psychology, 95,* 1134–1144. http://dx.doi.org/10.1037/a0020754

Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review, 14,* 490–495. http://dx.doi.org/10.5465/amr.1989.4308371

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17,* 601–617. http://dx.doi.org/10.1177/014920639101700305

*Wilson, K. S., & Baumann, H. M. (2015). Capturing a more complete view of employees' lives outside of work: The introduction and development of new interrole conflict constructs. *Personnel Psychology, 68,* 235–282. http://dx.doi.org/10.1111/peps.12080

*Zhang, Y., Waldman, D. A., Han, Y., & Li, X. (2015). Paradoxical leader behaviors in people management: Antecedents and consequences. *Academy of Management Journal, 58,* 538–566. http://dx.doi.org/10.5465/amj.2012.0995

(*Appendices follow*)

Appendix A

Detailed Instructions for Anderson and Gerbing's (1991) Approach

Please read the instructions very carefully. The questions are unique to survey measurement development and require detailed attention.

Research projects in the management field often use survey items to measure work concepts, such as work motivation, job satisfaction, and employee stress. When writing survey items, management researchers must take great care to ensure that the items do a good job of measuring the concepts of interest (e.g., that an item intended to measure work motivation really seems to capture that concept well). The goal of this study is to assess survey items used in the management literature.

Your job in this survey is to take each item in the left column and decide which concept it seems to best represent.

On the next few pages you will see lists of items next to three boxes. Each box contains a term and corresponding definition.

For each item, drag and drop the item to the box that it best matches.

Please pay close attention to each individual item as you decide which term and definition it best matches.

Before beginning the survey, below is an example to help guide your understanding of the survey.

Your job in this survey is to take each item in the left column and decide which concept it seems to best represent. Let's use the three concepts below as an example:

Work Motivation: The effort expended in relation to work.

Job Satisfaction: The enjoyment of work and job tasks.

Work Location: The location in which work is done.

Based on the terms and definitions above, an item that does a good matching ***Work Motivation: The effort expended in relation to work*** might be "I work hard in my job," because it speaks to a certain effort level at work. An item that also does a good job matching this term and definition might be, "I often feel lazy at the office," because it also speaks to a certain effort level at work. So you would drag and drop items like those to the box associated with ***Work Motivation: The effort expended in relation to work***. Please note that some of the items on the survey will focus on high levels of a given concept (like the "I work hard" item), whereas others will focus on low levels of a given concept (like the "I often feel lazy" item)—both can capture the concept of expending effort equally well.

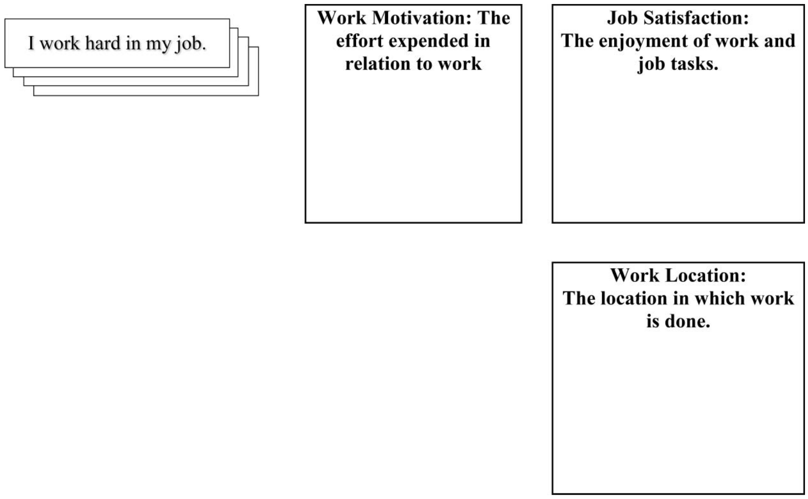
An item that does a good matching ***Job Satisfaction: The enjoyment of work and job tasks*** might be "I like coming to work," because it speaks to a certain enjoyment level at work. An item that also does a good job matching this term and definition might be "I think my job is boring," because it also speaks to a certain enjoyment level at work. So you would drag and drop items like those to the box associated with ***Job Satisfaction: The enjoyment of work and job tasks.***

What about an item like "I work in a city?" That item doesn't seem to have much to do with ***Work Motivation: The effort expended in relation to work***, so you wouldn't drag it to that box. That item also doesn't seem to have much to do with ***Job Satisfaction: The enjoyment of work and job tasks***, so you wouldn't drag it to that box either. It does seem to match ***Work Location: The location in which work is done***, so that would be the box you would drag that item to.

Please note that some of the items on the survey will seemingly match more than one term and definition. However, your job is to determine which definition the item best matches.

LET'S PRACTICE!

There are nine items in the stack below. Using the example above, drag and drop each item into the box with the term and definition it best matches.



I work hard in my job.

Work Motivation: The effort expended in relation to work

Job Satisfaction: The enjoyment of work and job tasks.

Work Location: The location in which work is done.

It is time to begin the real survey. On the next few pages you will be asked to take each item in the left column and decide which concept it seems to best represent. Also, please note that there will be a few questions that check how closely you're paying attention. Be sure to respond to these questions based on their directions.

Note. If participants drug and dropped any of the items into the inappropriate box, they were shown this error message: "Uh oh! Looks like you got something wrong. Please reread the directions and check your answers." The other eight items included: "I like coming to work," "I work in a city," "I think my job is boring," "I work in a tall building," "I often feel lazy at the office," "My work tasks are fun," "I lack energy when working," and "I work in a basement."

(*Appendices continue*)

Appendix B

Detailed Instructions for Hinkin and Tracey's (1999) Approach

Please read the instructions very carefully. The questions are unique to survey measurement development and require detailed attention.

Research projects in the management field often use survey items to measure work concepts, such as work motivation, job satisfaction, and employee stress. When writing survey items, management researchers must take great care to ensure that the items do a good job of measuring the concepts of interest (e.g., that an item intended to measure work motivation really seems to capture that concept well). The goal of this study is to assess survey items used in the management literature.

Your job in this survey is to assess the degree to which each item listed matches the statement provided.

On the next few pages you will see a **bolded** statement, followed by several survey items. For each item, you will rate the degree to which it matches the **bolded** statement. The items will repeat themselves on three consecutive pages, but the **bolded** statements will change. Again, simply rate the degree to which each item matches the **bolded** statement on that page.

Not all of the items will match the **bolded** statement. Therefore, please pay close attention to each individual question as you decide whether it matches the **bolded** statement.

Before beginning the survey, below is an example to help guide your understanding of the survey.

The survey asks you to judge how well a survey item matches particular statements, which will be presented to you in **bold**. You will make that judgment using this response scale:

1	**2**	**3**	**4**	**5**	**6**	**7**
Item does an	Item does a	Item does a	Item does	Item does a	Item does a	Item does an
EXTREMELY	**VERY**	**SOMEWHAT**	an	**SOMEWHAT**	**VERY**	**EXTREMELY**
BAD	**BAD**	**BAD**	**ADEQUATE**	**GOOD**	**GOOD**	**GOOD**
job of measuring	job of measuring	job of measuring	job of measuring	job of measuring	job of measuring	job of measuring
the **bolded** concept	the **bolded** concept	the **bolded** concept	the **bolded** concept	the **bolded** concept	the **bolded** concept	the **bolded** concept
provided above	provided above	provided above	provided above	provided above	provided above	provided above

For example, let's say the statement is: ***Work Motivation: The effort expended in relation to work.***

Since this statement refers to effort, an item that does a good job matching this statement might be, "I work hard in my job," because it speaks to a certain effort level at work. An item that also does a ***good*** job matching this statement might be, "I often feel lazy at the office," because it also speaks to a certain effort level at work. In contrast, an item that does a ***bad*** job matching this statement might be, "I work in a city," because it has very little to do with the effort level at work. Please note that some of the items on the survey will focus on high levels of a given concept (like the "I work hard" item), whereas others will focus on low levels of a given concept (like the "I often feel lazy" item). Both can capture the concept of expending effort equally well.

LET'S PRACTICE!

Using the example above, please rate the following three items on how well each does matching our concept, ***Work Motivation: The effort expended in relation to work.***

1. I work hard in my job.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
2. I work in a basement.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
3. I lack energy when working.	**1**	**2**	**3**	**4**	**5**	**6**	**7**

It is time to begin the real survey. On the next few pages you will be asked to evaluate how well items match a specific concept and statement.

Also, please note that there will be a few questions that check how closely you're paying attention. Be sure to respond to these questions based on their directions.

Note. If participants answered the three practice items inappropriately (i.e., anything other than a 5 through 7 for the first and third items and anything other than a 1 through 3 for the second item), they were shown this error message: "Uh oh! Looks like you got something wrong. Please reread the directions and check your answers."

Received May 4, 2017
Revision received February 11, 2019
Accepted February 12, 2019 ■