Journal of Business and Psychology (2022) 37:639–673
https://doi.org/10.1007/s10869-022-09813-3

ORIGINAL PAPER

Normalizing the Use of Single-Item Measures: Validation of the Single-Item Compendium for Organizational Psychology

Russell A. Matthews[1] · **Laura Pineault**[2] · **Yeong-Hyun Hong**[1]

Accepted: 25 March 2022 / Published online: 14 April 2022

Abstract

The application of single-item measures has the potential to help applied researchers address conceptual, methodological, and empirical challenges. Based on a large-scale evidence-based approach, we empirically examined the degree to which various constructs in the organizational sciences can be reliably and validly assessed with a single item. In study 1, across 91 selected constructs, 71.4% of the single-item measures demonstrated *strong* if not *very strong* definitional correspondence (as a measure of content validity). In study 2, based on a heterogeneous sample of working adults, we demonstrate that the majority of single-item measures examined demonstrated little to no comprehension or usability concerns. Study 3 provides evidence for the reliability of the proposed single-item measures based on test–retest reliabilities across the three temporal conditions (1 day, 2 weeks, 1 month). In study 4, we examined issues of construct and criterion validity using a multi-trait, multi-method approach. Collectively, 75 of the 91 focal measures demonstrated *very good* or *extensive* validity, evidencing moderate to high content validity, no usability concerns, moderate to high test–retest reliability, and extensive criterion validity. Finally, in study 5, we empirically examined the argument that only conceptually narrow constructs can be reliably and validly assessed with single-item measures. Results suggest that there is no relationship between subject matter expert evaluations of construct breadth and reliability and validity evidence collected across the first four studies. Beyond providing an off-the-shelf compendium of validated single-item measures, we abstract our validation steps providing a roadmap to replicate and build upon. Limitations and future directions are discussed.

Keywords Single-item measure · Validity · Reliability · Organizational sciences

In the organizational sciences, it is seemingly an urban legend that to validly assess psychological constructs, one *must* use multi-item measures (e.g., Allen et al. 2022; Boyd et al. 2005). While it has long been argued that the use of single-item measures should not constitute a "fatal flaw" (e.g., Wanous et al. 1997), resistance to their applicability continues (Boyd et al. 2005; Singh 2003). To be clear, as discussed in the larger literature (e.g., Cheah et al. 2018; Fuchs and Diamantopoulos 2009), there are constructs where single-item measures are likely *not appropriate*. For example, constructs that incorporate multi-dimensional definitions, or for constructs where items may be interpreted differently in heterogeneous samples, are not good candidates for assessment with single-item measures. More specifically, multiple items may be required to ensure respondents' assessment of theoretically essential aspects of a conceptually complex construct; in such cases, a single-item measure may lead respondents to make an ambiguous and general interpretation of the construct without their considerations of all aspects of the construct (Fuchs and Diamantopoulos 2009). From a psychometric perspective, more items can average out random error across items, which improves (certain types of) reliability (Sarstedt and Wilczynski 2009), which can allow for increased measurement accuracy (Peter 1979), with the potential for greater construct validity (Wanous et al. 1997). And more practically, multi-item measures allow for different options when dealing with issues like

✉ Russell A. Matthews
ramatthews2@ua.edu

Laura Pineault
laura.pineault@wayne.edu

Yeong-Hyun Hong
yhong18@crimson.ua.edu

[1] University of Alabama, 361 Stadium Drive, Box 870225, Tuscaloosa, AL 35487, USA

[2] Wayne State University, 5057 Woodward Ave., Suite 8402.3, Detroit, MI 48202, USA

missing data (Cheah et al. 2018). The straw man argument then is that because multi-item measures have these potential advantages, single-item measures are somehow inherently deficient. This seemingly knee-jerk reaction that all single-item measures are in some way not valid or imply a weak research design is counterproductive and may inadvertently limit advancements in the organizational sciences.

Outside the organizational sciences, including fields like epidemiology, nursing, and political sciences, the use of single-item measures is much more accepted. As a result, a broad body of interdisciplinary work has detailed the validity and utility of single-item measures (e.g., Ang and Eisend 2018; Cheah et al. 2018; De Vries et al. 2016). While we return to this issue in more detail shortly, single-item measures have two primary advantages. First, they minimize practical concerns (e.g., survey length, participant fatigue, response and retention rates; Fuchs and Diamantopoulos 2009) when applied to various types of data collection efforts. Second, they often have fewer issues related to contamination and redundancy (compared to multi-item measures; Fisher et al. 2016; Wanous et al. 1997), wherein measurement contamination can result in spurious relationships between constructs or suppress the observed relationship because the construct is not correctly assessed (Guion 1965). Put another way, asking tangentially related items (i.e., contamination) or asking the basic same item over and over again (i.e., redundancy) simply to ensure a multi-item measure is available can result in unintended consequences (Arnulf et al. 2018; Boyle 1991; Maul 2017).

Admittedly, a major hurdle related to substantiating the use of single-item measures in the organizational sciences is that many were developed with limited evidence to support their reliability and validity (Fuchs and Diamantopoulos 2009). Taking up the call to action by Allen et al., (2022, p. 1), who noted that "now more than ever, it is essential to ensure that single-item measures are valid and reliable," we take a large-scale evidence-based approach to examine the degree to which constructs common to the organizational sciences can be assessed *reliably* and *validly* with single-item measures. That is, rather than subjective evaluations regarding the appropriateness of single-item measures (Singh 2003), a systematic evaluation of their applicability in the organizational sciences not only is warranted but also serves as a means to advance the field by leveraging their inherent benefits (Allen et al. 2022).

Our research makes several contributions to the literature. First, we present an evidence-based approach towards the underlying utility of single-item measures to help highlight the degree to which constructs in the organizational sciences lend themselves to this measurement approach (i.e., reduce subjective evaluations of what is an appropriate measurement approach). Put simply, just as we do not believe single-item measures represent an inherent fatal flaw, we are not arguing that all constructs are amenable to this assessment approach. That is, for example, Fuchs and Diamantopoulos (2009) present a conceptual checklist (e.g., construct concreteness and complexity, semantic redundancy, desired precision) for evaluating if a construct should be assessed with a single-item measure. Building on previous literature then, we seek to leverage data to better guide this ongoing discussion. For example, we investigate whether the reliability and validity of single-item measures are related to the degree to which constructs are conceptually *narrow* or *complex*.

Second, based on our systematic approach to understanding the reliability and validity of single-item measures, we provide an evidence-based template for future research, within and beyond the organizational sciences, to examine other constructs. Finally, and more practically, we provide scholars and practitioners with a series of measures they can confidently use to draw valid inferences in their own research. We see this as particularly advantageous given single-item measures may serve as a potentially proactive solution in addressing the research-practice gap (Lapierre et al., 2018). Moreover, there are a host of ongoing calls encouraging organizational scientists to engage in more cross-disciplinary research. Having a compendium of validated single-item measures may help organizational scientists more effectively align their methodological approaches with and facilitate collaborations across disciplines where single-items are more commonly applied.

Our program of research includes five studies. In study 1, we examine the content validity of 91 single-item measures. In study 2, we explore usability and comprehension concerns (Gehlbach and Brinkworth 2011; Rossiter 2002). In study 3, we examine the test–retest reliability of the single-item measures across three distinct time-unit conditions (i.e., 1 day, 2 weeks, 1 month; Dormann and Van de Ven 2014). In study 4, we examine issues of construct and criterion validity. Finally, in study 5, we empirically examine the underlying argument (e.g., Fuchs & Diamantopoulos 2009) that the reliability and validity of single-item measures decrease as the conceptual breadth of a construct increases (i.e., that single-item measures should only be used for conceptually *narrow* versus *complex* constructs). First though, we provide a brief primer on some of the benefits of single-item measures.

Benefits of Single-Item Measures

While various excellent summaries exist (e.g., Allen et al. 2022; Cheah et al. 2018; Fuchs and Diamantopoulos 2009), holistically, single-item measures have two overarching advantages. First, they help minimize significant practical concerns. For example, leveraging single-item measures

within a larger program of research can help mitigate issues of respondent burden, survey length, and item repetition (Drolet and Morrison 2001; Rogelberg and Stanton 2007). These advantages are particularly salient in more resource intensive data collection efforts (e.g., diary studies, experience sampling designs, multi-source designs, international surveillance projects). In turn, efficiently using survey space may allow researchers to include additional relevant constructs (i.e., address issues of model deficiency) while still balancing against concerns like increased respondent burden (Wanous et al. 1997).

Thus, the judicious use of single-item measures can help researchers retain respondents who may not be interested in engaging in a lengthy survey, preventing non-response biases and survey breakoff (Göritz 2014). Their use can also reduce the cognitive demands placed on respondents compared with multi-item measures, helping to minimize insufficient response effort patterns (e.g., straight-lining) and missing variables (Fuchs and Diamantopoulos 2009). Furthermore, the application of single-item measures can help researchers adapt their assessment approach to new research contexts (Cheah et al. 2018; Nagy 2002).

Second, beyond salient practical concerns, single-item measures may demonstrate fewer issues with criteria contamination while still being construct valid (Drolet and Morrison 2001; Fisher et al. 2016; Wanous and Hudy 2001). Single-item measures tend to present a revised version of the construct definition or include several content-relevant examples. This is done to proactively protect against issues of criterion deficiency (Wanous and Hudy 2001) but has the added benefit of ensuring construct-irrelevant characteristics (i.e., criterion contamination; Scarpello and Campbell 1983) are not assessed. For example, Drolet and Morrison (2001) empirically demonstrated that while multi-item measures result in strong estimates of internal consistency, beyond a well-developed focal item, additional items add very little explanatory power in terms of capturing and understanding the underlying construct but can meaningfully increase error term correlations across items.

While long-standing concerns exist around their validity and reliability (Cronbach and Meehl 1955; Nunnally 1978; Nunnally and Bernstein 1994), there is no theoretical reason to argue that all single-item measures are "deficient." Rather, just as with multi-item measures, it is incumbent on researchers to provide validity and reliability information for proposed single-item measures. And while validity and reliability information for single-item measures may differ from those commonly reported for multi-item measures (Allen et al. 2022), we echo the argument that it is possible to develop valid and reliable measures for many constructs in the organizational sciences (Bergkvist and Rossiter 2007; Fisher et al. 2016; Spörrle and Bekk 2014).

Construct Selection and Item Development

Initially, we considered over 200 constructs for inclusion based on a systematic review of prestigious journals within the organizational sciences (e.g., *Journal of Applied Psychology*, *Journal of Management*, *Personnel Psychology*, *Journal of Business & Psychology*). We applied an iterative review process, based on a series of inclusion and exclusion criteria, to reduce the number of constructs. First, to increase the overall utility of the resulting measures, we emphasized constructs that have received systematic attention in the literature. Next, we identified and refined, as needed, operational definitions for each focal construct to account for nuanced differences and changes in construct definitions over time and across subdisciplines (Slaney and Garcia 2015), excluding constructs with inconsistent conceptual definitions. In turn, we only considered constructs where valid multi-item measures existed (to be able to examine issues of construct validity). We also generally excluded constructs where validated single-item measures already existed (e.g., Fisher et al. 2016; Gilbert and Kelloway 2014; Yarkoni 2010).

We should note that *construct complexity* (Fuchs and Diamantopoulos 2009), or the degree to which a construct is conceptually narrow (simple) versus broad (complex), was *not* used as an initial inclusion criterion. The general argument is that as construct breadth (i.e., conceptual complexity) increases, it is necessary to include more items in a measure to ensure adequate sampling of the target construct (i.e., to maximize content validity; Thurstone 1947), wherein increased representation of the conceptual space increases predictive validity (Ones and Viswesvaran 1996). Thus, it has been recommended that only conceptually narrow (i.e., simple) constructs be assessed with single-item measures (Fuchs and Diamantopoulos 2009). To our knowledge though, there is no established procedure for evaluating the degree to which a construct is *broad* versus *narrow*.[1] Rather, researchers are generally encouraged to rely on their "professional judgement" (Gehlbach and Brinkworth 2011, p. 383) when deciding on the number of items needed to represent a construct (Hinkin 1998). Given the relative subjectivity involved in evaluating construct breadth, we did not include this as an initial criterion; rather, this is an issue we return to in study 5.

[1] We should note that definitional and terminological confusion surrounds the concepts of *complexity* and *concreteness*, with different authors approaching and defining terms in inconsistent ways. For example, referring to a construct as "broad" or "abstract" could be meant to refer to how large the construct space is (e.g., a "complex" personality trait), to denote that a construct is multi-dimensional in nature (e.g., job satisfaction), or to denote that a concept is not grounded in sensory-motor information (e.g., role ambiguity; Borghi et al. 2017).

With that in mind though, two additional considerations played a key role in selecting focal constructs and the item writing process. First, to increase the utility of the items across different research methodologies, we sought to ensure that items were interpretable across different recall windows (e.g., daily recall, the past month); we wanted to validate items that were versatile in terms of temporal cadence (i.e., study 3). Thus, items were phrased such that, in future research, scholars can more confidently adjust the recall window to match theoretically relevant temporal processes under consideration (Dormann and Van de Ven 2014). Second, common instruction stems and response formats (e.g., Likert and frequency) were applied to standardize the survey administration process. These steps should help scholars effectively manage the cognitive load on participants while also partially addressing issues of common method variance by leveraging two different response scales (Podsakoff et al. 2003).

Thus, based on each construct definition, we developed an item that either presented content-relevant examples and/or presented a revised version of the definition. Consistent with established recommendations (e.g., Gehlbach and Brinkworth 2011), prior to our formal data collection efforts, 20 psychology and management Ph.D. graduate students, with past training and experiences with various scale development efforts, evaluated construct definitions for accuracy and provided suggested changes to the single-item measures (e.g., language complexity, item clarity). Relevant changes were made as needed. Based on this collected process, 91 constructs were identified for further evaluation and validation (see Table 1).

Study 1

Demonstrating content validity is "an initial step toward construct validation by all studies which use new, modified, or previous unexamined measures" (Schriesheim et al. 1993, p. 385) and was the focus of study 1. Among suggested ways to examine content validity (e.g., Anderson and Gerbing 1991; Hinkin and Tracey 1999), we assessed *definitional correspondence*, "the degree to which a scale's items correspond to the construct's definition" (Colquitt et al. 2019, p. 1243).[2] Per Hinkin and Tracey (1999), the only requirement for making content validity judgments 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). Thus, similar to Colquitt et al. (2019), we leveraged a large sample of naïve raters (working adults) given they are ideal for establishing estimates of content validity because they are representative of samples where the measures might be administrated. Because each item either included several content-relevant examples based explicitly on the construct definition and/or presented a revised version of the construct definition itself, per classification standards developed by Colquitt et al. (2019), we predict that the majority of single-item measures will demonstrate content validity as evidenced by naïve raters' evaluations of definitional correspondence.[3]

Hypothesis 1 *Single-item measures demonstrate acceptable levels of definitional correspondence (i.e., definitional correspondence estimates* ≥ 0.60*).*

Method

Participants and Procedure

Participants were recruited via Prolific.co. Only employed US residents with a 98% or higher approval rating were permitted to participate based on pre-established screeners via Prolific; respondents were paid $2.85. Consistent with established recommendations (Curran 2016; Huang et al. 2012), respondents who failed to correctly complete at least three of four effortful responding questions were excluded. We allowed respondents to miss one attention check item given respondents may mistakenly miss one item but still, generally, be attentive (Huang et al. 2012; McGonagle et al. 2016).

Of the 610 respondents, we excluded 19 for not meeting inclusion criteria (i.e., not currently working) and another 30 for failing multiple attention checks or for nonsensical responses to open-ended questions. The analysis sample ($N = 561$) was 50.7% female, primarily Caucasian (77.2%) with an average age of 34.7 years ($SD = 10.61$) and

[2] We recognize that content validity is also often conceptualized in terms of the degree to which a construct is accurately captured by the item(s) included in a given measure. While we follow the example set forth by Colquitt et al. (2019) in terms of examining definitional correspondence as an indicator of content validity, in other literatures, our approach might be evaluated as an examination of face validity. As discussed by Allen et al., (2022, p. 1) though, "just as for multi-item measures, it is critically important for single-item measures to demonstrate face validity," wherein face validity can be defined as the "clarity or relevance" of an item or measure.

[3] Colquitt et al. (2019) provide overall criteria with five levels (Table 5 in their study). A definitional correspondence estimate of .91 and above is considered *very strong*, .87 to .90 is *strong*, .84 to .86 is *moderate*, .60 to .83 is *weak*, and .59 and below as *lack of* definitional correspondence. To be clear, we are not arguing that a definitional correspondence estimate of .60 to .83, which Colquitt et al. again define as *weak*, is necessarily acceptable. Consistent with our overarching argument for triangulating the validity of single-item measures, the goal is to ensure content validity is demonstrated first and foremost and then evaluated against other pieces of psychometric evidence relative to the needs of a specific program of research.

Table 1 Single-item measures and their content validity, usability concerns, and reliability estimates

Construct	Single-item (reading level)	Definitional correspondence (S1)			Usability concerns (S2)	Reliability (ICC/*r*) (S3)		
		SI	MI	*t*(df)		1-day	2-weeks	1-month
Abusive supervision — active-aggressive	My supervisor was abusive, saying or doing things to me that were openly hostile, harsh, or insulting. (10.4)[a, 1]	.91	.85	6.42** (118)	2.6% DNA	.47/.47	.59/.59	.54/.54
Affective commitment	I felt emotionally attached to my organization. (6.3)[a, 1]	.91	.88	1.77 (106)	.5% DNA	.73/.73	.78/.78	.69/.69
Authoritarian leadership	My supervisor asserted absolute control and authority over the people he/she supervises, demanding obedience from them. (5.4)[a, 1]	.93	.80	8.78** (125)	3.1% DNA	.70/.71	.51/.51	.49/.49
Autonomy climate	My organization designs jobs in ways which give employees flexibility about how and when to enact their work. (11.7)[c, 1]	.94	.57	21.12** (139)	1.3% DNA .5% CD	.54/.54	.61/.61	.46/.46
Bureaucracy	My organization has a lot of bureaucracy; every decision has to be approved by someone higher up. (9.9)[c, 1]	.90	.79	9.07** (142)	1.0% DNA .3% CQ .3% CD .3% CR	.64/.64	.71/.71	.55/.55
Career satisfaction	I am satisfied with my career. (4.4)[c, 1]	.82	.82	−.09 (102)	1.5% DNA .3% CD	.87/.81	.84/.84	.79/.76
Climate for civility	My organization expects employees to treat one another with respect. (11.9)[c, 1]	.92	.85	5.09** (111)	.8% DNA .3% CR	.57/.58	.73/.58	.49/.49
Cognitive demands	My work was very cognitively demanding. (10.3)[a, 1]	.92	.77	9.28** (122)	.5% DNA 5% CQ .3% CD	.55/.55	.60/.60	.65/.65
Competitive goals	Where I work, each person competed against other employees to achieve their own goals. (8.4)[a, 1]	.84	.62	10.66** (122)	2.6% DNA	.63/.63	.57/.57	.48/.48
Competitive orientation	I thought of competition with others at work as a way to enhance my development and to demonstrate my self-worth. (10.5)a, 1	.86	.71	7.86** (116)	3.6% DNA	.56/.56	.47/.47	.47/.47
Continuance commitment — high sacrifices	I have felt like I stay with my organization mainly because I would lose out on too much if I left. (7.7)[a, 1]	.87	.85	1.94 (124)	1.8% DNA .5% CQ .5% CD	.62/.62	.55/.55	.59/.59
Continuance commitment-low alternatives	I have felt like I stay with my organization mainly because there is nowhere else for me to get a similar or better job. (10.4)[a, 1]	.91	.93	6.47** (123)	2.0% DNA .5% CD	.59/.61	.65/.65	.50/.51
Cooperative goals	Where I work, people worked together cooperatively to achieve work goals. (9.0)[a, 1]	.91	.76	8.26** (122)	.5% DNA	.61/.62	.51/.52	.35/.35
Cooperative orientation	I viewed others at work as partners, wherein I was willing to work with them to achieve common goals. (7.3)[a, 1]	.90	.76	8.60** (136)	.8% DNA	.56/.56	.47/.47	.44/.44
Coworker trust	I felt like I could trust my coworkers. (2.2)[a, 1]	.92	.79	8.33** (113)	1.0% DNA .3% CD	.68/.67	.60/.60	.63/.63
Daily work hassles	Did you have to deal with minor, but nonetheless irritating, issues at work? (11.1)[b, 2]	.86	.69	7.30** (126)	.3% DNA	.57/.56	.64/.64	.58/.58

Table 1 (continued)

Construct	Single-item (reading level)	Definitional correspondence (S1)			Usability concerns (S2)	Reliability (ICC/*r*) (S3)		
		SI	MI	*t*(df)		1-day	2-weeks	1-month
Deep acting	When I was interacting with others at work I really tried to feel the emotions I was expected to show. (9.3)[a, 1]	.80	.82	− 1.15 (108)	.3% DNA 2.3% CQ .5% CD	.46/.48	.26/.26	.16/.16
Demands-abilities job fit	I felt like my personal abilities and skills were a good fit with the requirements and demands of my job. (9.3)[a, 1]	.89	.88	.86 (112)	–	.52/.53	.66/.65	.51/.51
Distributive justice	My supervisor made sure that opportunities and rewards were distributed fairly. (13.3)[a, 1]	.93	.75	9.19** (112)	5.4% DNA	.51/.51	.64/.64	.50/.50
Efficiency climate	My organization places a lot of emphasis on employee efficiency and productivity at work. (14.3)[c, 1]	.91	.66	10.26** (119)	.8% DNA .5% CD	.33/.34	.48/.48	.55/.55
Emotional demands	Was your work emotionally demanding? (10.7)[b, 2]	.85	.80	3.77** (119)	–	.71/.71	.74/.74	.59/.59
Emotional fatigue	Did you suffer from emotional fatigue related to your work, wherein you had extreme emotional tiredness and/or an inability to feel or show emotions at the end or the work day? (16.7)[b, 2]	.88	.81	4.92** (112)	.3% DNA	.71/.71	.68/.68	.69/.70
Extrinsic motivation	At work, I was motivated by the opportunities and rewards I could receive regardless of the extent to which I enjoyed the work I did. (12.0)[a, 1]	.88	.78	4.27** (132)	1.0% DNA .3% CQ .3% CD	.47/.46	.50/.51	.23/.28
Face-time orientation	Where I work, if you want to advance and get ahead, you have to put in a lot of "face time." (5.5)[c, 1]	.90	.81	5.87** (116)	2.6% DNA .5% CQ .5% CD	.65/.65	.49/.49	.33/.33
Family authenticity	The time, energy, and attention I gave to my FAMILY was consistent with my life values and priorities (11.0)[a, 1]	.91	.86	4.33** (148)	1.3% DNA	.69/.69	.66/.66	.46/.46
Family motivation	I do this job because I want to support and take care of my family. (5.2)[c, 1]	.94	.88	6.35** (121)	2.6% DNA .5% CD	.73/.73	.72/.72	.60/.62
Formalization Climate	My organization places a lot of emphasis on formal rules and procedures. (9.7)[c, 1]	.93	.57	21.97** (122)	.5% DNA .5% CD .3% CR	.58/.54	.65/.66	.55/.56
Goal-focused leadership	My supervisor encouraged employees' to work towards achieving goals communicated by the organization. (17.6)[a, 1]	.88	.72	9.07** (127)	3.3% DNA	.58/.41	.58/.58	.49/.49
Informational justice	My supervisor explained decisions that affected me, and my work, in a thorough and timely way. (9.8)[a, 1]	.90	.76	8.55** (119)	3.3% DNA .3% CQ	.56/.56	.66/.67	.53/.53
Innovation climate	My organization places a lot of emphasis on encouraging and supporting new ideas and innovative approaches. (14.9)[c, 1]	.92	.78	10.19** (119)	.8% DNA .3% CD .3% CR	.64/.64	.61/.60	.72/.72
Interpersonal justice	My supervisor was generally respectful and polite when discussing work related issues with me. (13.4)[a, 1]	. 91	.83	5.46** (115)	2.8% DNA	.59/.60	.71/.71	.41/.42

Table 1 (continued)

Construct	Single-item (reading level)	Definitional correspondence (S1)			Usability concerns (S2)	Reliability (ICC/*r*) (S3)		
		SI	MI	*t*(df)		1-day	2-weeks	1-month
Intrinsic motivation	I did the work I did because it was inherently interesting and satisfying. (9.4)[a, 1]	.93	.83	6.96** (138)	.3% DNA	.72/.73	.69/.70	.75/.60
Intuitive decision-making style	At work, I made decisions quickly, relying on intuition (feelings) to do so. (7.6)[a, 1]	.93	.85	7.97** (129)	.3% DNA	.56/.56	.54/.54	.32/.31
Job insecurity	I felt like there was a good chance I could lose my job. (1.2)[a, 1]	.90	.70	12.27** (114)	1.0% DNA	.40/.40	.73/.73	.72/.73
Job self-efficacy	I felt like I had the skills and abilities to perform well in my job. (5.2)[a, 1]	.92	.89	2.45* (115)	–	.55/.57	.77/.63	.54/.54
Leader avoidant conflict behaviors	My supervisor avoided conflicts rather than directly addressing it. (14.1)[a, 1]	.84	.78	4.85** (128)	2.8% DNA	.35/.35	.43/.43	.61/.62
Leader collaborative conflict behaviors	When there were conflicts at work, my supervisor encouraged people to engage in constructive negotiations and collaborative problem solving. (15.4)[a, 1]	.93	.80	8.37** (123)	6.4% DNA .3% CD	.62/.61	.52/.62	.24/.34
Leader dominating conflict behaviors	When there were conflicts at work, my supervisor encouraged people to compete and "win the battle" (win the conflict). (9.8)[a,1]	.90	.79	7.93** (120)	5.4% DNA .3% CD	.69/.69	.43/.43	.28/.28
Leadership self-identity	I see myself as a leader. (2.4)[c, 1]	.90	.83	6.78** (115)	–	.85/.85	.78/.78	.70/.70
Learning goal orientation	At work, I was motivated by a desire to develop myself by acquiring new skills, mastering new situations, and improving my competence. (13.9)[a, 1]	.92	.78	9.75** (117)	.5% DNA	.59/.60	.63/.63	.50/.49
Loneliness	Did you feel lonely at work? (2.2)[b, 2]	.88	.81	6.74** (140)	.3% CQ	.68/.69	.65/.65	.70/.70
Managerial responsibility stress	Did you feel stressed because you were responsible for supervising or managing other people at work? (11)[b, 2]	.86	.76	6.56** (157)	4.8% DNA	.61/.61	.66/.66	.67/.67
Meaning	I found my work to be very meaningful. (3.7)[a, 1]	.86	.76	6.56** (157)	.3% DNA	.74/.74	.78/.78	.66/.66
Meeting effectiveness	The meetings I was involved in at work were effective. (4.8) a, 1	.74	.72	.67 (114)	3.1% DNA	.51/.51	.63/.63	.50/.50
Mental fatigue	Did you suffer from mental fatigue, wherein you had extreme mental tiredness and an inability to think or concentrate at the end or the work day? (13.4)[b, 2]	.90	.91	5.59** (116)	–	.68/.58	.70/.70	.65/.65
Needs-supplies job fit	I felt like the things that I needed from my job were fulfilled by what my job offered me, financially, socially, and/or psychologically. (11.9)[a, 1]	.94	.81	8.92** (123)	.8% DNA	.71/.72	.63/.63	.52/.52
Negative effort-reward imbalance	I invested more in my job than I received in return. (4.7)[a, 1]	.88	.81	6.75** (125)	.3% DNA .3% CD	.69/.69	.64/.64	.57/.57

Table 1 (continued)

Construct	Single-item (reading level)	Definitional correspondence (S1)			Usability concerns (S2)	Reliability (ICC/*r*) (S3)		
		SI	MI	*t*(df)		1-day	2-weeks	1-month
Normative commitment	I felt like I stayed with my organization mainly because I felt that I "ought to." (6.8)[a, 1]	.88	.80	3.46** (111)	1.3% DNA .3% CD	.69/.69	.53/.53	.34/.34
Organizational politics	There was a lot of "organizational politics" where I work. (8.3)[a, 1]	.79	.80	−.25 (112)	.5% DNA .5% CQ .3% CD	.70/.70	.61/.61	.56/.56
Organizational reputation	The organization I work for has a good reputation. (8.8)[c, 1]	.83	.91	1.64 (122)	.3% DNA	.47/.75	.75/.75	.67/.67
Ostracism	Did you feel like you were ignored or excluded by others at work? (5.9)[b, 2]	.88	.77	8.08** (131)	.5% DNA	.71/.71	.73/.73	.62/.62
Perceived contract breach	I feel like my organizations has broken a lot of its "promises" to me. (7.5)[c, 1]	.84	.62	8.39** (116)	1.8% DNA .3% CR	.52/.68	.68/.68	.41/.42
Perceived leader effectiveness	My supervisor was an effective leader, he/she helped me perform my job well. (7.5)[a, 1]	.92	.78	9.03** (117)	2.8% DNA	.71/.71	.70/.70	.57/.57
Perceived organizational support	I felt like my organization was very supportive of me. (8.3)[a, 1]	.87	.84	1.51 (117)	1.0% DNA	.67/.67	.65/.65	.69/.71
Perceived overqualification	I felt like I was overqualified for the job I have. (4.7)[a, 1]	.90	.83	5.70** (134)	.8% DNA	.74/.74	.74/.74	.67/.67
Performance goal orientation	At work, I was motivated by a desire to show myself, and others around me, that I was competent and able to do my job effectively. (12.2)[a, 1]	.90	.74	9.39** (123)	–	.53/.53	.45/.45	.04/.04
Person-organization fit	I felt like I "fit" with my organization. (5.2)[a, 1]	.79	.85	−4.25** (134)	1.0% DNA	.66/.66	.73/.73	.62/.61
Perspective-taking	I made an effort to take the perspective of other people at work, actively seeking out opportunities to understand their viewpoints. (12.8)[a, 1]	.92	.86	5.47** (159)	1.0% DNA	.39/.39	.43/.44	.48/.48
Pessimism of organizational change	I was doubtful that any program or company effort to solve problems where I work would actually make a difference. (10.5)[a, 1]	.86	.83	2.09* (128)	2.0% DNA 1.3% CQ .5% CD	.29/.30	.44/.44	.30/.30
Physical fatigue	Did you suffer from physical fatigue, wherein you had extreme physical tiredness and an inability to engage in physical activity at the end or the work day? (16)[b, 2]	.87	.80	4.71** (118)	.3% DNA	.62/.62	.69/.70	.59/.60
Preference for group work	If given the choice, I prefer working in a team than alone at work. (5.0)[c, 1]	.94	.89	5.59** (129)	.5% DNA	.73/.72	.77/.77	.61/.60
Procedural justice	My supervisor made sure that his/her decisions were made fairly and ethically based on accurate information and unbiased procedures. (14.6)[a, 1]	.87	.77	5.77** (118)	3.3% DNA .3% CD	.58/.58	.66/.67	.50/.50
Prosocial identity	I see myself as caring and generous. (5.6)[c, 1]	.95	.89	5.95** (121)	–	.64/.64	.64/.65	.59/.60

Table 1 (continued)

Construct	Single-item (reading level)	Definitional correspondence (S1)			Usability concerns (S2)	Reliability (ICC/*r*) (S3)		
		SI	MI	*t*(df)		1-day	2-weeks	1-month
Prosocial motivation	At work, I was motivated by my desire to help (benefit) other people. (8.0)[a, 1]	.91	.86	3.85** (108)	.3% DNA	.56/.56	.64/.64	.54/.64
Quality of group experience	General speaking, the people I work with all get along with one another. (7.6)[c, 1]	.85	.88	− 2.00* (117)	.3% DNA	.54/.54	.54/.55	.46/.46
Quantitative workload	Did you feel like you had a heavy workload, with lots to do? (4.4)[b, 2]	.87	.73	9.63** (121)	–	.67/.58	.69/.68	.63/.63
Rational decision-making style	At work, I made decisions thoughtfully, approaching them in a rationale way based on facts. (8.3)[a, 1]	.92	.86	5.31** (126)	.3% DNA .3% CD	.55/.56	.50/.50	.40/.40
Relationship conflict	Was there interpersonal conflict among the people you work with? (8.7)[b, 2]	.81	.76	3.21** (135)	.8% DNA	.46/.47	.66/.66	.61/.61
Resources	I had the resources I needed to do my job effectively (e.g., equipment, training, information, technical support). (12.5)[a, 1]	.93	.82	7.04** (119)	.3% DNA .3% CD	.61/.61	.44/.44	.43/.44
Role conflict	Did you experience conflicting expectations for what you need to do at work? (9.8)[b,2]	.85	.70	7.51** (104)	–	.52/.52	.57/.57	.58/.58
Rumination (negative)	Did you find yourself thinking about bad things that happened at work? (5.8)[b, 2]	.82	.83	− .07 (114)	.3% DNA .3% CQ	.60/.61	.68/.68	.55/.56
Self-initiated work breaks	Did you take a short break when you needed one while at work? (3.6)[b, 2]	.72	.84	− 5.03** (113)	–	.62/.62	.64/.64	.48/.48
Subjective monotony	Did you find your job boring? (2.2)[b, 2]	.84	.84	− .14 (121)	–	.76/.77	.71/.71	.76/.76
Subjective stress	Did you find your job stressful? (2.2)[b, 2]	.77	.70	3.47** (105)	–	.70/.70	.73/.73	.79/.78
Supervisor competence	My supervisor was highly competent at his/her job. (7.5)[a, 1]	.94	.81	9.51** (122)	2.8% DNA .5% CD	.69/.69	.74/.73	.57/.57
Supervisor warmth	My supervisor was good-natured, warm, and sincere. (9.0)[a, 1]	.93	.83	7.02** (117)	2.6% DNA .5% CD	.65/.66	.74/.74	.54/.54
Surface acting	When I was interacting with others at work, I often felt like I had to hide (fake) what I was really feeling. (8.5)[a, 1]	.89	.85	3.30** (117)	.5% DNA	.57/.57	.62/.62	.62/.62
Task conflict	People I worked with often disagreed about how we should work together to accomplish our work. (9.0)[a, 1]	.84	.79	3.77** (118)	.8% DNA	.35/.35	.42/.42	.46/.46
Task interdependence	I had to work with other employees in my organization to get my work done effectively. (9.8)[a, 1]	.84	.54	7.63** (114)	.5% DNA	.52/.52	.50/.50	.42/.41
Team self-management	My team and I are responsible for deciding how we are going to do our work, not a manager or supervisor. (10.0)[c, 1]	.89	.88	1.01 (139)	2.3% DNA .3% CD	.54/.54	.60/.61	.54/.54
Team workload sharing	Everyone I worked with did their fair share of the work. (0.9)[a, 1]	.91	.82	6.45** (115)	.8% DNA	.66/.66	.63/.63	.49/.48

Table 1 (continued)

Construct	Single-item (reading level)	Definitional correspondence (S1)			Usability concerns (S2)	Reliability (ICC/*r*) (S3)		
		SI	MI	*t*(df)		1-day	2-weeks	1-month
Time pressure	Did you feel like you did not have enough time to complete your work? (4.6)[b, 2]	.92	.80	7.83** (120)	–	.54/.54	.42/.42	.42/.42
Training climate	My organization places a lot of emphasis on developing employee skills. (12.3)[c, 1]	.88	.83	3.20** (113)	1.0% DNA	.44/.67	.79/.79	.52/.52
Trust in supervisor	I felt I could trust my supervisor. (3.9)[a, 1]	.91	.81	5.62** (125)	2.6% DNA	.78/.78	.82/.82	.49/.49
Unnecessary illegitimate tasks	I had to carry out tasks at work that were unnecessary, or could be done easier if things were better organized. (9.4)[a, 1]	.92	.72	11.51** (121)	.5% DNA	.55/.55	.67/.67	.65/.65
Unreasonable illegitimate tasks	I had to do unreasonable things at work that fell outside of my job responsibilities and should be done by someone else. (11.2)[a, 1]	.92	.73	10.72** (113)	.5% DNA	.74/.74	.71/.71	.44/.44
Welfare climate	My organization values and cares for its employees. (9.6)[c, 1]	.92	.72	13.06** (116)	.5% DNA	.57/.80	.79/.79	.63/.63
Work authenticity	The time, energy, and attention I gave to my WORK was consistent with my life values and priorities (9.7)[a, 1]	.92	.85	4.68** (128)	.3% DNA	.58/.58	.56/.57	.45/.45
Work frustration	Did you feel frustrated while at work? (3.6)[b, 2]	.79	.68	5.72** (122)	–	.71/.71	.74/.71	.72/.72
Work hypercompetitive	I was extremely competitive at work. (8.3)[c, 1]	.91	.85	4.53** (130)	.8% DNA	.70/.70	.60/.60	.57/.57
Work pressure	Did you feel like you were working under a lot of pressure at work? (5.4)[b, 2]	.81	.78	1.53 (123)	–	.63/.62	.72/.72	.81/.81

S study, *CQ* confusing question, *CD* cannot decide, *CR* cannot remember, *DNA* does not apply, *ICC* inter-class correlation

[a]"Thinking about the past [insert recall window]"

[b]"Thinking about the past [insert recall window], how often…"

[c]"Please answer the following."

[1]1–5 disagree–agree

[2]1–5 frequency

*$p < .05$; **$p < .01$

organizational tenure of 5.43 years (*SD* = 5.51). On average, respondents worked 37.09 h/week (*SD* = 10.61).

As per Colquitt et al. (2019), respondents were asked to evaluate how well a given item "matched" the construct's conceptual definition (i.e., *definitional correspondence*) based on a 5-point scale (1 = *Not at all* to 5 = *To a very great extent*). For a given construct, respondents rated the proposed single-item measure. Respondents also rated items from a previously published multi-item measure of the focal construct. Definitional correspondence ratings for existing multi-item measures were collected for diagnostic purposes.

We administered all items for a given construct in a randomized block. To manage response fatigue, we randomized the presentation of the focal constructs, wherein only twenty constructs were presented to each respondent. Respondents completed a sample content evaluation exercise (with corrective feedback) prior to engaging in the larger assessment.

Measures

Table 1 reports on all single-item measures. Construct definitions are reported in the Online Supplemental Materials as are the multi-item reference measures.

Results

On average, each construct had 123 (*SD* = 10.34) definitional correspondence estimates. To evaluate content validity as a function of definitional correspondence, Colquitt et al. (2019) provide overall criteria with five levels (Table 5 in their study). To apply their criteria, we divided the definitional correspondence estimate for each single-item measure by the number of response options (*a* = 5). For the multi-item measures, per Colquitt et al., we averaged definitional correspondence estimates across the items in the scale divided by the number of response options (*a* = 5). Table 1 reports definitional correspondence estimates; 39 measures demonstrated *very strong* definitional correspondence (0.91 and above), 26 demonstrated *strong* estimates (0.87 to 0.90), fourteen demonstrated *moderate* estimates (0.84 to 0.86), and twelve demonstrated *weak* estimates (0.60 to 0.83). None demonstrated a *lack of* definitional correspondence (0.59 and below) suggesting that Hypothesis 1 is supported; the single-item measures demonstrate definitional correspondence.

Supplemental Analyses

In the case of the previously published multi-item measures, only three demonstrated *very strong* definitional correspondence, eight demonstrated *strong* estimates, fifteen demonstrated *moderate* estimates, 62 demonstrated *weak* estimates, and three demonstrated a *lack of* definitional correspondence. In turn, we compared definitional correspondence scores for the two approaches based on a paired-sample *t*-test for each construct. Across the 91 constructs, the multi-item construct demonstrated higher definitional correspondence 3.3% of the time. There was no significant observed difference for 14.3% of the constructs. Definitional correspondence was significantly higher for single-item measures for the remaining 82.4% of constructs.

Study 2

Collectively, 71.4% of the single-item measures demonstrated *strong* or *very strong* content validity; while content validity is only a piece of the validity puzzle, and definitional correspondence is just one facet of content validity, evidence from study 1 is encouraging. That said, a primary concern with single-item measures is that while the selected item itself may be content valid, that item may not fully capture the entirety of the construct (i.e., have restricted content adequacy; Hinkin and Tracey 1999). To overcome this issue, single-item measures tend to be longer (i.e., in terms of word count), present more content-relevant examples within the item, and/or present a revised version of the construct definition. We applied all of these approaches to minimize issues of construct deficiency in our single-item measures (Fuchs and Diamantopoulos 2009). However, doing so runs the inherent risk that the resulting measures are meaningfully complex or difficult for respondents to understand, process, and respond to in a thoughtful way (Peter 1979). The trade-off then is that in addressing issues of construct coverage, single-item measures may engender respondent usability concerns.[4]

Readability is "the ease with which a reader can read and understand text" (Oakland and Lane 2004, p. 244) and is an item characteristic shown to influence reliability (Tourangeau et al. 2020). Several factors contribute to text readability (Dubay 2004), including, for example, the number of words per sentence (Flesch 1948; Kincaid et al. 1975). In practice, readability is often estimated based on the Flesch-Kincaid method, wherein readability scores are reflective of the US reading grade level required to effectively understand a statement, phrase, or passage (Crossley et al. 2008); specific to our study then, a higher readability score for a given single-item would be indicative of that item being

[4] Please note that presenting multiple examples in an item does not mean an item is necessarily *double-barreled*. Double-barreled questions ask about two distinct (i.e., divergent) "attitudinal" phenomena wherein respondents provide only one answer (Olson 2008). Using the word "and" in an item does not inherently make it a double-barreled item. However, in the spirit of study 2, the use of multiple examples (conjuncts) may increase the complexity and impact the interpretability of single-item measures (Olson 2008).

more complex or difficult to understand. Psychometricians have encouraged scholars to use readability indices to help simplify psychological assessments and to make the content easier to understand. Doing so promotes item application across contexts and populations (e.g., different age cohorts, educational levels), and reduces construct irrelevant variance, such as *g* (Fowler 1995). We suggest that readability provides an avenue to index and explain potential respondents' *usability concerns*. Respondents may experience usability concerns when answering single-item measures especially because, in trying to ensure content adequacy (Hinkin and Tracey 1999), a given item may be too difficult for respondents to understand and respond to. Thus, first, we examined if respondents reported systematic response difficulties with the measures. In turn, drawing on existing work (Tourangeau et al. 2020), we would predict that items with higher readability scores should result in more usability concerns.

Hypothesis 2 *More complex single-item measures (indexed based on reading level) have more participant usability concerns.*

Method

Participants and Procedure

Participants were recruited by 47 students from a university in Southern USA who distributed a standardized email invitation to working adults they personally knew; students received nominal course credit. To ensure a heterogeneous sample, the only inclusion criteria were that respondents be 18 years of age or older and working at least part-time. Respondents were asked to complete the survey (online), which included the single-item measures and basic demographics. If respondents did not feel a given item applied to them, or felt unable to answer a given item, they were asked to use a drop-down menu and indicate the reason.

A total of 421 respondents completed the survey; 29 indicated they were not working and were excluded. The final sample ($N = 392$) was 65.6% female, with an average age of 38.21 years ($SD = 14.63$) and organizational tenure of 7.87 years ($SD = 9.07$). On average, respondents worked 39.45 h/week ($SD = 12.47$), wherein 87.0% reported having a direct supervisor, and 39.8% reported supervising other employees. The sample was ethnically diverse with 69.4% Caucasian, 17.3% African American, 2.6% Hispanic, 2.6% Asian, and 4.1% identifying as multi-racial. Approximately 34.7% of the sample reported working something other than a regular daytime shift. While 17.3% of respondents did not report working in a work group, other respondents reported working in a variety of work group contexts, such that 26.5% reported working in one work group; another 26.5% reported not only working in one primary group, but also having a secondary workgroup; and 29.3% reported working in more than one workgroup.

Measures

The same 91 single-item measures (Table 1) were used. Participants were first instructed, "If you can answer the question, use the options below" and were presented with the appropriate frequency or Likert scale. In turn, participants were instructed, "If you can't answer a question, please use the drop-down to indicate why." Four possible options were presented: the item *did not apply* to them, the question *did not make sense* to them, they *could not decide how to answer*, and they *could not remember* (respondents could only select one option). Participants were asked to consider the past month when responding.

Results

Table 1 provides a summary of how often respondents endorsed a reason they could not answer a given item and the estimated reading level for each item ($M = 8.64$, $SD = 3.76$). We applied the Flesch-Kincaid method for calculating grade level within Microsoft Word, which is based on US reading grade-level averages (Crossley et al. 2008).

Means, standard deviations, and bivariate correlations are reported in the Online Supplemental Materials. Given sample heterogeneity, it is not surprising that when respondents felt they could not respond to an item, the most common reason was that it did not apply to them; items referencing a supervisor had higher endorsement of *does not apply* (e.g., *leader dominating conflict behaviors* — 6.4%). More relevant to usability concerns are the remaining response options. Only two items had more than 1% of the sample indicate that a question was confusing (i.e., *deep acting* — 2.3%; *pessimism of organizational change* — 1.3%). Less than 1% of the sample endorsed *could not decide how to answer* or *could not remember* for any given item. Respondents' reports indicated a minimal level of usability concerns across the single-item measures, providing confidence in the comprehensibility of the measures.

Specific to Hypothesis 2, reading level was not correlated with endorsement rates for *did not make sense* ($r = 0.03$, $p = 0.78$), *could not decide how to answer* ($r = 0.11$, $p = 0.29$), or *could not remember* ($r = 0.17$, $p = 0.11$). Given respondents could only select one option, but all three represent "usability" concerns, we created a composite endorsement score — this overall score was also unrelated to reading level ($r = 0.10$, $p = 0.37$). Hypothesis 2 was not supported, likely because so few usability concerns were observed.

Supplemental Analyses

Given approximately 65% of our sample reported having at least a bachelor's degree, we examined if education level influenced usability concerns such that the sample was split. Group 1 ($n = 135$) reported an education level as *some college, but no degree*, or less. Group 2 ($n = 253$) reported an education level of *Associate's* or higher. For each respondent, the number of usability concerns reported was summed across the 91 items (the response option *does not apply* was excluded). The independent samples *t*-test was not statistically significant [$t(386) = 0.253$, $p = 0.80$]; there was no difference in usability concerns reported by group 1 ($M = 0.19$; $SD = 0.81$) compared to group 2 ($M = 0.17$; $SD = 0.65$).

Study 3

Once again, study 1 provides initial evidence for the content validity of the single-item measures and study 2 suggests that, other than items that are potentially population specific (e.g., items that apply only to people with a supervisor), usability concerns are limited at best. With this foundation, we conducted study 3 to assess the reliability of the proposed single-item measures. As noted though, given the nature of single-item measures, the most commonly used index of reliability, internal consistency, is not possible. Thus, scholars have advocated the use of test–retest reliabilities in the context of single-item measures (Spörrle and Bekk 2014; Wanous and Hudy 2001). To provide a robust understanding, we examined the test–retest reliabilities of the single-item measures across three temporal conditions, over a 1-day lag, a 2-week lag, and a 1-month lag such that these temporal lags may represent conceptually distinct time units (Dormann and Van de Ven 2014).

The most common procedure to index test–retest reliability is Pearson's *r*. However, there is limited guidance defining what is an "acceptable" Pearson *r* test–retest reliability. As such, some have argued evaluating test–retest reliability using intra-class correlation coefficient (ICC; wherein "0" is indicative of no reliability and "1" is indicative of excellent reliability) given evaluative ICC criteria have been established. Thus, we calculated Pearson's *r* and ICC test–retest reliabilities (Koo and Li 2016) but leveraged existing ICC guidelines for evaluating single-item test–retest reliabilities.[5]

Hypothesis 3 *Single-item measures demonstrate acceptable levels of test–retest reliability (i.e., ICC ≥ 0.40).*

[5] Cicchetti (1994) suggests that ICC values greater than .74 indicate excellent reliability, between .60 and .74 indicate good reliability, between .40 and .59 indicate fair reliability, and below .40 indicate poor reliability.

Method

Participants and Procedure

Participants were again recruited via Prolific (study 1 respondents were excluded from participating), wherein only employed US residents with a 95% approval rating or higher were permitted to participate. However, building on study 2 results, we screened respondents to have a supervisor by using pre-established screeners within the platform to help ensure items were applicable to all respondents. Participants were informed that this was a two-part study. In part 1, based on a between-person experimental design, respondents were randomly assigned to one of the three time-unit conditions (1-day lag, 2-week lag, and 1-month lag). Respondents were invited back to complete part 2 of the study based on the time-unit condition they were assigned. To match the temporal lag for a given condition, conditions were formatted the same except for the recall window that respondents were asked to consider. Respondents were paid $1.50 and $1.25 for participating in the first and second survey, respectively. Respondents who failed to correctly complete at least three of four effortful responding questions (in either survey) were excluded (Huang et al. 2012).

A total of 628 individuals responded to the initial survey; 45 were excluded for not meeting inclusion criteria (i.e., not currently working or not having a supervisor). Another 32 were excluded for failing multiple attention checks, for nonsensical responses to open-ended questions, or for not consenting to participate in the second study. Thus, we retained an analysis sample of 551 respondents at time 1 (condition 1 = 180; condition 2 = 191; condition 3 = 180).

Condition 1 (1-day lag) had a time 2 response rate of 82.4% ($N = 155$). We excluded eight respondents for either failing to finish the survey or for missing multiple attention checks. Also, given condition 1 was meant to replicate daily-diary type research, another 20 respondents who indicated they did not work on either day the survey was administered were excluded. This resulted in an analysis sample of 127 of which 47.2% was female, with an average age of 32.42 years ($SD = 9.50$) and organizational tenure of 4.49 years ($SD = 5.17$). On average, respondents worked 38.13 h/week ($SD = 9.59$).

Condition 2 (2-week lag) had a time 2 response rate of 82.7% ($N = 158$). We excluded three respondents for either failing to finish the survey or missing multiple attention checks. Another eight respondents were excluded because they reported a major job change between survey administrations. This resulted in an analysis sample of 147 of which 53.7% was female, with an average age of 33.80 years ($SD = 9.90$) and organizational tenure of 5.45 years ($SD = 4.88$). On average, respondents worked 37.93 h/week ($SD = 10.20$).

Condition 3 (1-month lag) had a time 2 response rate of 70.2% ($N = 132$). We excluded four respondents for either failing to finish the survey or missing multiple attention checks. Another six respondents were excluded because they reported a major job change between survey administrations. This resulted in an analysis sample of 122 of which 52.5% was female, with an average age of 35.08 years ($SD = 10.28$) and organizational tenure of 5.14 years ($SD = 5.39$). On average, respondents worked 39.37 h/week ($SD = 9.58$).

No demographic differences were observed across the three conditions for gender, age, organizational tenure, or hours worked.

Results

We report means, standard deviations, and bivariate correlations for all items in the Online Supplemental Materials for each condition. ICC estimates (2-way mixed-effects model with absolute agreement with type set as single measurement), by condition, are reported in Table 1, followed by Pearson's *r* test–retest correlations. As Pearson's *r* test–retest correlations are the more common way researchers have established test–retest reliability, we first examined the degree to which ICC reliability estimates correlated with Pearson's *r* reliability estimates. In condition 1, the approaches correlated at 0.89 ($p < 0.001$); in condition 2, they correlated at 0.98 ($p < 0.001$) and at 0.98 ($p < 0.001$) in condition 3 as well. Results suggest that the two approaches result in similar estimates of reliability; again though, the primary benefit of ICC-based test–retest reliability estimates is the ability to apply existing evaluative criteria (Cicchetti 1994).

In condition 1, four measures (4.4%) demonstrated excellent reliability, 44 (48.4%) demonstrated good reliability, 38 (41.8%) demonstrated fair reliability, and five (5.5%) demonstrated poor reliability. In condition 2, ten measures (11.0%) demonstrated excellent reliability, 53 (58.2%) demonstrated good reliability, 27 (29.7%) demonstrated fair reliability, and one (1.1%) demonstrated poor reliability. And in condition 3, five measures (5.5%) demonstrated excellent reliability, 25 (27.8%) demonstrated good reliability, 51 (56.0%) demonstrated fair reliability, and ten (11.0%) demonstrated poor reliability. In support of Hypothesis 3, across the three conditions, 94.1% of the time the focal constructs demonstrated an ICC test–retest reliability above 0.40.

Supplemental Analyses

Once again, recognizing that no single piece of psychometric information is sufficient, we designed our research to triangulate in on the construct validity of the proposed single-item measures. Interestingly then, there was no relationship between content validity (Table 1) and ICC test–retest reliability in any of three temporal lags ($r_{\text{Condition 1}} = 0.08$, $p = 0.44$; $r_{\text{Condition 2}} = -0.01$, $p = 0.92$; $r_{\text{Condition 3}} = -0.15$, $p = 0.16$).

Study 4

Collectively, we have demonstrated that the single-item measures show acceptable levels of content validity, limited usability concerns due to issues of cognitive complexity, and acceptable levels of test–retest reliability across three temporal conditions. Thus, in study 4, we first focus on investigating the degree to which the proposed measures demonstrate construct validity and then turn our attention to the broader issue of criterion validity.

Construct validity is defined as the "the correspondence between a construct and a measure taken as evidence of the construct" (Edwards 2003, p. 329). Within the single-item measurement literature, scholars often seek to establish construct validity by demonstrating that the proposed single-item measure "loads" with items from an existing multi-item measure of the same construct (e.g., Fisher et al. 2016), wherein it is implicitly assumed that the multi-item measure is a "valid" measure of the construct. We apply this approach to establish the degree to which the single-item measures "tap into" the underlying conceptual construct.

Similar to Fisher et al. (2016), we conducted a series of confirmatory factor analyses (CFA), one for each construct, wherein the single-item measure for a given focal construct as well as items from a previously published multi-item measure for that construct (i.e., the same measures used in study 1) were loaded on to the same latent factor. A standardized factor loading can be considered as the extent to which the single-item measure correlates with the corresponding conceptual latent construct (McDonald 1999), evidencing construct validity; if a single-item measure failed to load significantly on the latent factor (wherein the model, otherwise, demonstrated good fit), that would suggest that the item does not tap into the same construct, suggesting a lack of construct validity.

A benefit of using CFA models to examine construct validity is that we can leverage existing recommendations related to the interpretation of factor loadings (i.e., higher factor loadings for the single-item measure is indicative of higher construct validity for that item; Fisher et al. 2016). Within the multi-item measurement literature, the generally recommended minimum for interpreting factor loadings is 0.32; that is, albeit relatively *poor*, factor loadings between 0.32 and 0.44 demonstrate at least minimal construct validity (Comrey and Lee 1992; Tabachnick and Fidell 2013). By extension, Comrey and Lee have argued that factor loadings of 0.45 to 0.54 are *fair*, loadings of 0.55 to 0.62 are *good*, loadings of 0.63 to 0.70 are *very good*, and loadings 0.71 and greater are *excellent*.

Hypothesis 4 *As an indicator of construct validity, single-item measures load significantly on the underlying latent factor with factor loadings* ≥ 0.32.[6]

Until this point, our focus has been on establishing "internal" psychometric characteristics of the proposed single-item measures, that is, do they *accurately* and *precisely* measure the construct of interest (Hughes 2018). This focus is intentional given that demonstrating evidence of psychometric accuracy and precision (i.e., content validity; response process validity evidence; test–retest reliability evidence; structural evidence) is considered "an initial step toward construct validation" (Schriesheim et al. 1993, p. 385). However, we now turn to the broader issues of criterion validity, which is established by demonstrating that the measure of an intended construct (i.e., a proposed single-item measure) is related to the measure of some other alternative constructs that it should be related to, based on theoretical or conceptual arguments. The issue at hand is that, based on classical test theory, the magnitude of a criterion correlation cannot exceed the product of the reliability indices (Schmitt 1996). And given the pervasive argument in the literature that single-item measures are somehow unreliable, it has been argued that single-item measures have lower criterion related validity compared to multi-item measures of the same construct (for a relevant discussion, see Ziegler et al. 2014).

Beyond that, the concern over content adequacy (Hinkin and Tracey 1999) is one of the most pervasive arguments against the use of single-item measures. That is, by using several indicators, multi-item measures are, conceptually, better able to represent underlying construct space (Guion 1965; Thurstone 1947). Specific to criterion validity then, the implication is that multi-item measures, by being "more reliable" and by better capturing the construct, may demonstrate more accurate and systematic relationships with conceptual relevant alternative construct compared to single-item measures (Cheah et al. 2018). Put another way, if content adequacy and reliability are such pervasive problems across single-item measures, it may be unreasonable to expect single-item measures to achieve the same level of criterion validity as multi-item measures. To date, studies examining the issue have generally demonstrated that single-item measures do evidence criterion-related validity (e.g., Cheah et al. 2018; Spörrle and Bekk 2014). However, the majority of single-item measurement studies published to date have focused on validating a single focal construct. As such, there is a potential for a "file-drawer" problem to exist such that scholars may have an overconfidence in the approach because measures that do not evidence criterion validity may not have been published (Allen et al. 2022).

Our focus on such a broad spectrum of constructs provides a unique opportunity to understand the degree to which single-item measures demonstrate criterion validity, especially relative to multi-item measures of the same construct. Again then, returning to our guiding premise, while we think many constructs in the organizational sciences can be assessed with single-item measures, that does not mean that all single-item measures will demonstrate the same level of criterion validity as a reference multi-item measure of the same construct. Beyond that, if single-item measures do in fact demonstrate systematically lower criterion validity, that would place a meaningful boundary condition on their general utility to organizational scientists, irrespective of how often they are used in other fields.

To examine this issue, we adopt Campbell and Fiske's (1959) multi-trait multi-method (MTMM) matrix approach such that we examine the degree to which criterion validity relationships differ as a function of method of assessment (i.e., single- vs multi-item measures of the construct). For example, we would expect that a measure of abusive supervision should be related to a measure of interpersonal justice (Lian et al. 2012). The question is, within a MTMM approach, does the single-item measure of abusive supervision correlate at approximately the same level with the criterion construct, interpersonal justice, as the multi-item measure of abusive supervision? To examine this issue more concretely, we apply Raykov's (2011) procedure for interval estimation of convergent validity coefficients.

Research Question: *To what extent do single-item measures demonstrate criterion validity as compared to multi-item measures of the same construct?*

Method

Participants and Procedure

Given the potential for a loss in response quality during long surveys (e.g., Bowling et al. 2021; Meade and Craig 2012), it was not feasible to present all 91 single-item measures as well as the corresponding multi-item measures (a total of 474 items, plus demographic questions) to all participants. Thus,

[6] To be clear, similar to Hypothesis 1, we are not advocating that single-item measures demonstrating "poor" construct validity (i.e., loadings of .32 to .44) are necessarily valid. We set this minimum based on accepted practices in the larger scale development literature. As noted by Allen et al. (2022) though, standards applied to validating single-item measures may be different than those used for multi-item measures. Thus, depending on the construct under consideration, setting more stringent minimums might be prudent. It should also be recognized that using different multi-item measures of the same focal construct might result in different construct validity estimates for a given single-item measures. Again then, single-item construct validity evidence must be interpreted relative to other pieces of validity evidence including content validity as well as the psychometric characteristics of the comparative multi-item measure.

participants were randomly assigned to complete approximately 1/3 of the single-item measures and 2/7 of the multi-item measures. We used random assignment (instead of yoking a given single-item measure to the multi-item measure for a given construct) to examine issues of construct validity (i.e., Hypothesis 4) and criterion validity (i.e., Research Question) within and across constructs. Because of this, a large sample was needed to ensure adequate power for any possible combination of constructs.

Following study 2 protocols, participants were recruited by 201 college students (there was no overlap in student recruiters from study 2). Recruiters were solicited from 23 classes across 19 geographically dispersed universities in the USA. Initially, 1444 respondents were recruited; we excluded 71 who did not meet inclusion criteria and another 52 who did not finish the survey. The final sample ($N = 1321$) was 59.4% female, with an average age of 35.5 years ($SD = 15.59$) and organizational tenure of 6.5 years ($SD = 8.04$). On average, respondents worked 40.6 h/week ($SD = 11.97$), wherein 37.2% reported supervising other employees. The sample was ethnically diverse with 58.4% Caucasian, 13.1% African American, 9.5% Hispanic, 6.4% Asian, and 6.0% identifying as multi-racial (another 4.2% declined to respond). A third of the sample (34.1%) reported working something other than a regular daytime shift. While 18.8% of respondents did not report working in a work group, other respondents reported working in a variety of work group contexts, such that 24.1% reported working in one work group, another 34.5% reported working in one primary group but also having a secondary workgroup, and 22.3% reported working in more than one workgroup.

Measures

Table 1 reports on all single-item measures, and the same multi-item measures from study 1 were used (see Online Supplemental Materials for more information). Multi-item measures with negatively worded items were reverse coded. We report internal consistency reliability estimates for the multi-item measures in Table 2.

Results

Means, standard deviations, and bivariate correlations for all construct, for both approaches, are reported in the Online Supplemental Materials.

Construct Validity

To examine issues of construct validity (Hypothesis 4), we conducted 91 CFA models (*Mplus*; Muthén and Muthén, 2018), one for each construct. All items from the multi-item reference construct were set to load on a single factor, as was the single-item measure for that construct. Given respondents were randomly assigned to complete a subset of single-item and multi-item measures, listwise deletion was applied to ensure respondents completed both the single-item and multi-item measure of the focal construct.

Table 2 includes a summary of CFA results. Overall, the models demonstrated acceptable fit (Hu and Bentler 1999); the poorest fitting models were for *autonomy climate* and *perceived contract breach.*[7] In support of Hypothesis 4, the proposed single-item measure significantly loaded ($\beta \geq 0.32$ and $p < 0.05$; Table 2) on the underlying latent factor (in the correct direction) for 84 of the 91 constructs. Three single-item measures (*work hypercompetitive*, *performance goal orientation*, and *face-time orientation*) loaded significantly, but their standardized factor loadings were less than 0.32. Another three (*efficiency climate*, *extrinsic motivation*, and *task interdependence*) did not load significantly, and one (*normative commitment*) loaded significantly, but in the opposite direction. Based on established criteria (Comrey and Lee 1992), 33 single-item measures demonstrated *excellent* construct validity (36.3%), 17 demonstrated *very good* construct validity (18.7%), 13 demonstrated *good* construct validity (14.3%), 17 demonstrated *fair* construct validity (18.7%), four demonstrated *poor* construct validity (4.4%), and seven failed to demonstrate construct validity (7.7%).[8]

Criterion Validity

To examine issues of criterion validity, we adopted a MTMM matrix approach (within *Mplus*, based on syntax from Raykov 2011), wherein we included both single- and multi-item measures of the focal construct (e.g., *abusive supervision*) and the criterion construct (e.g., *interpersonal justice*; again, both the single- and multi-item measures were included, albeit we focus on the multi-item measure of the criterion construct). In turn, based on Raykov, we estimated the confidence interval of the correlation difference between

[7] These two multi-item measures included negatively worded items. *Autonomy climate* ($\chi^2(5) = 83.04$, $p < .001$, CFI = .83, SRMR = .07) and *perceived contract breach* ($\chi^2(5) = 110.48$, $p < .001$, CFI = .89, SRMR = .08) both continued to demonstrate poor fit when the respective single-item measure was excluded and the CFA model re-estimated; the poor fit seems to be a function of the multi-item measure, not because of single-item measure.

[8] Another way to evaluate construct validity is to examine the bivariate correlation between the two measurement approaches. As such, the bivariate correlations between the single-item and multi-item reference measure, for each construct, are reported in Table 2. The average construct validity correlation across the 91 constructs was .58 ($SD = .19$). Interestingly, across the 91 constructs, single-item CFA factor loadings correlated at .97 ($p < .001$) with the observed bivariate correlations between the single-item and multi-item construct measures suggesting that the two approaches for establishing construct validity are effectively equivalent.

Table 2 Single-item measures and their construct and criterion-related validity, study 4

Construct (MI alpha)	*r*	*N*	χ^2 (DF)	CFI	SRMR	Single-item		Criterion validity			
						Loading	Com	Construct	SI-MI	MI-MI	95% CI
Abusive supervision — active-aggressive (.94)	.78**	134	29.72** (9)	.98	.02	**.80**	.63	Interpersonal justice	−.65**	−.70**	−.14, .04
								Per. leader effectiveness	−.52**	−.58**	−.20, .08
								Climate for civility	−.33**	.31**	−.12, .16
Affective commitment (.89)	.70**	147	15.99** (2)	.96	.04	**.72**	.52	**Person-organizational fit**	.38**	.62**	.11, .36
								Perceived organizational support	.32**	.43**	−.05, .26
								Intrinsic Motivation	.25**	.45**	.10, .38
Authoritarian leadership (.78)	.61**	146	7.89* (2)	.97	.03	**.68**	.46	Abusive supervision	.41**	.35**	−.25, .12
								Bureaucracy	.28**	.53**	.08, .40
								Unreasonable tasks	.33**	.24**	−.30, .13
Autonomy climate (.76)	.44**	140	66.56** (9)	.73	.09	**.49**	.24	*Procedural justice*	.47**	.29**	−.37, −.01
								Innovation climate	.51**	.41**	−.31, .11
								Work frustration	−.34**	−.23**	−.09, .30
Bureaucracy (.90)	.49**	132	12.29 (9)	.99	.03	**.50**	.25	*Organizational politics*	.54**	.34**	−.37, −.03
								Autonomy climate	−.50**	−.75**	−.37, −.11
								Unreasonable tasks	.41**	.37**	−.22, .15
Career satisfaction (.91)	.60**	123	27.42** (9)	.96	.04	**.62**	.39	*Meaning*	.65**	.45**	−.33, −.05
								Needs-supply fit	.62**	.51**	−.25, .04
								Affective commitment	.50**	.49**	−.17, .15
Climate for civility (.88)	.54**	146	12.69* (5)	.98	.03	**.54**	.30	Welfare climate	.55**	.53**	−.18, .13
								Quality of group experience	.43**	.80**	.23, .50
								Organizational politics	−.42**	−.50**	−.23, .08
Cognitive demands (.77)	.48**	124	8.44 (5)	.98	.03	**.54**	.29	Quantitate workload	.44**	.47**	−.14, .20
								Work pressure	.47**	.27**	−.38, −.02
								Emotional demands	.42**	.54**	−.03, .28
Competitive goals (.89)	.53**	142	11.27* (5)	.99	.03	**.57**	.32	**Organizational politics**	.52**	.75**	.10, .36
								Role conflict	.38**	.66**	.14, .43
								Task conflict	.37**	.44**	−.14, .30
Competitive orientation (.80)	.47**	148	37.48** (14)	.93	.05	**.54**	.29	**Performance goal orientation**	.25**	.57**	.16, .48
								Work hypercompetitive	.19*	.45**	06, .46
								Competitive goals	.19*	.24*	−.17, .27
Continuance commitment—high sacrifices (.78)	.45**	142	.59 (2)	1.00	.01	**.49**	.24	CC Low alternatives	.33**	.46**	−.10, 35
								Normative commitment	.20**	.33**	−.07, .34
								Task conflict	.16*	.21*	−.15, .24
Continuance commitment-low alternatives (.84)	.66**	131	2.30 (2)	1.00	.01	**.71**	.50	Perceived contract breach	.36**	.35**	−.20, .17
								Affective commitment	−.36**	−.23**	−.04, .30
								Career satisfaction	−.35**	−.33**	−.22, .24
Cooperative goals (.88)	.58**	146	10.04 (5)	.99	.02	**.61**	.37	Coworker trust	.55**	.50**	−.19, .09
								Quality of group experience	.54**	.56**	−.12, .14
								Welfare climate	.53**	.41**	−.31, .06

Table 2 (continued)

Construct (MI alpha)	*r*	*N*	χ^2 (DF)	CFI	SRMR	Single-item		Criterion validity			
						Loading	Com	Construct	SI-MI	MI-MI	95% CI
Cooperative orientation (.87)	.39**	113	41.17** (20)	.94	.05	**.42**	.18	Cooperative goals	.51**	.52	−.21, .23
								Coworker trusts	.37**	.32**	−.23, .14
								Interpersonal justice	.29**	.41**	−.05, .29
Coworker trust (.86)	.76**	145	24.22** (5)	.95	.04	**.85**	.72	*Competitive goals*	−.56**	−.34**	.08, .36
								Procedural justice	.45**	.51**	−.08, .20
								Affective commitment	.41**	.29**	−.29, .04
Daily work hassles (.85)	.44**	134	12.66 (9)	.99	.03	**.48**	.23	Time pressure	.59**	.43**	−.37, .05
								Work frustration	.53**	.54**	−.15, .18
								Efficiency climate	−.42**	−.24**	.00, .36
Deep acting (.90)	.35**	124	2.39 (2)	1.00	.01	**.37**	.13	Emotional demands	.26**	.39**	−.06, .31
								Perspective taking	.26**	.34**	−.14, .30
								Performance goal orientation	.20**	.18	−.25, .20
Demands-abilities job fit (.86)	.65**	136	.06 (2)	1.00	.00	**.70**	.48	Needs-supply fit	.43**	.53**	−.06, .26
								Career satisfaction	.33**	.36**	−.16, .22
								Perceived org. support	.21**	.29**	−.10, .25
Distributive justice (.94)	.50**	129	13.13* (5)	.98	.02	**.51**	.26	*Leader effectiveness*	.64**	.35**	−.44, −.15
								Procedural justice	.59**	.58**	−.16, .14
								Perceived contract breach	−.53**	−.44**	−.07, .25
Efficiency climate (.84)	.03	132	5.44 (5)	1.00	.03	−.06	.00	Innovation climate	.41**	.38**	−.25, .20
								Goal-focused leadership	.36**	.38**	−.19, .24
								Unnecessary tasks	−.32**	−.42**	−.32, .11
Emotional demands (.88)	.77**	126	13.40* (5)	.98	.02	**.82**	.68	Emotional fatigue	.58**	.50**	−.26, .09
								Subjective stress	.58**	.51**	−.19, .06
								Rumination (negative)	.49**	.48**	−.15, .13
Emotional fatigue (.95)	.78**	111	20.18 (14)	.99	.02	**.79**	.63	Rumination (negative)	.58**	.69**	−.01, .24
								Subjective stress	.48**	.58**	−.05, .25
								Unreasonable tasks	.40**	.27**	−.29, .03
Extrinsic motivation (.75)	−.05	143	4.02 2)	.98	.03	−.09	.01	Distributive justice	.42**	.35**	−.28, .14
								Needs-supply fit	.41**	.28**	−.36, .10
								Leader effectiveness	.30**	.17*	−.35, .10
Face-time orientation (.72)	.19*	167	22.42** (9)	.93	.05	**.21**	.04	Organizational politics	.30**	.13	−.38, .03
								Bureaucracy	.06	.29**	.02, .45
								Task conflict	.24**	.23**	−.21, .20
Family authenticity (.91)	.74**	126	5.48 (2)	.99	.02	**.77**	.59	Resources	.38**	.46**	−.05, .23
								Time pressure	−.33**	−.43**	−.28, .08
								Welfare climate	.21**	.39**	.00, .37
Family motivation (.91)	.46**	95	23.16** (9)	.97	.05	**.42**	.18	CC high sacrifices	.28**	.26**	−.26, .21
								Emotional demands	.16	.22*	−.18, .30
								Prosocial motivation	.10	.39**	.08, .49

Table 2 (continued)

Construct (MI alpha)	*r*	*N*	χ^2 (DF)	CFI	SRMR	Single-item		Criterion validity			
						Loading	Com	Construct	SI-MI	MI-MI	95% CI
Formalization climate (.75)	.46**	140	7.55 (9)	1.00	.03	**.54**	.29	*Goal-focused leadership*	.29**	.05	−.44, −.03
								Abusive supervision	−.14*	−.22*	−.28, .11
								Bureaucracy	.10	.13	−.25, .30
Goal-focused leadership (.90)	.46**	147	8.41 (9)	1.00	.02	**.49**	.24	Leader collab. conflict	.60**	.64**	−.10, .17
								Perceived leader effectiveness	.59**	.65**	−.07, .20
								Innovation climate	.53**	.52**	−.19, .16
Informational justice (.91)	.61**	160	21.37* (9)	.98	.03	**.64**	.41	Goal-focused leadership	.55**	.64**	−.02, .22
								Leader warmth	.54**	.72**	.06, .29
								Meeting effectiveness	.49**	.47**	−.20, .16
Innovation climate (.91)	.54**	119	23.1 (14)	.98	.03	**.58**	.34	**Welfare climate**	.63**	.75**	.02, .24
								Organizational reputation	.50**	.54**	−.10, .18
								Resources	.47**	.45**	−.19, .17
Interpersonal justice (.91)	.73**	120	11.22* (5)	.99	.02	**.78**	.61	Informational justice	.72**	.62**	−.24, .02
								Abusive supervision	−.63**	−.70**	−.17, .04
								Perceived leader effectiveness	.60**	.68**	−.03, .20
Intrinsic motivation (.88)	.51**	154	2.24 (2)	1.00	.02	**.54**	.29	Career satisfaction	.54**	.51**	−.22, .16
								Needs-supply fit	.50**	.46**	−.18, .12
								Meaning	.48**	.53**	−.10, .20
Intuitive decision-making style (.81)	.49**	151	10.79 (9)	.99	.03	**.56**	.31	*Work pressure*	.26**	.05	−.40, −.01
								Unnecessary tasks	.25**	.34**	−.09, .26
								Organizational politics	.22**	.12	−.29, .10
Job insecurity (.61)	.50**	153	9.37** (2)	.98	.04	**.61**	.37	Competitive goals	.49**	.42**	−.23, .10
								Abusive supervision	.39**	.42**	−.11, .17
								Interpersonal justice	−.38**	−.26**	−.05, .30
Job self-efficacy (.90)	.53**	119	2.64 (2)	1.00	.02	**.55**	.31	**Demands-ability fit**	.45**	.73**	.13, .46
								Career satisfaction	.21*	.23*	−.21, .23
								Meaning	.18*	.25**	−.14, .27
Leader avoidant conflict behaviors (.85)	.62**	152	11.48* (5)	.98	.03	**.68**	.46	Leader effectiveness	−.46**	−.30**	−.03, .34
								Quality of group experience	−.32**	−.22*	−.10, .30
								Pessimism of org. change	.32**	.34**	−.15, .19
Leader collaborative conflict behaviors (.86)	.61**	142	11.36* (5)	.98	.03	**.66**	.43	Informational justice	.68**	.68**	−.11, .11
								Supervisor competence	.56**	.64**	−.06, .22
								Affective commitment	.43**	.49**	−.09, .20
Leader dominating conflict behaviors (.78)	.45**	147	3.58 (2)	.99	.02	**.51**	.26	Competitive goals	.30**	.36**	−.14, .26
								Interpersonal justice	−.29**	−.37**	−.28, .13
								Task conflict	.24**	.25**	−.22, .23

Table 2 (continued)

Construct (MI alpha)	*r*	*N*	χ^2 (DF)	CFI	SRMR	Single-item		Criterion validity			
						Loading	Com	Construct	SI-MI	MI-MI	95% CI
Leadership self-identity (.87)	.81**	118	4.47 (2)	.99	.02	**.88**	.77	Learning goal orientation	.36**	.36**	−.20, .20
								Preference for group work	.23*	.33*	−.27, .48
								Meaning	.23**	.29**	−.15, .29
Learning goal orientation (.87)	.48**	135	40.10** (14)	.95	.04	**.49**	.24	Intrinsic motivation	.49**	.52**	−.15, .19
								Innovation climate	.49**	.32**	−.36, .03
								Work authenticity	.35**	.32**	−.23, .16
Loneliness (.86)	.74**	147	3.19 (2)	1.00	.01	**.79**	.62	Rumination (negative)	.46**	.27**	−.40, .00
								Emotional fatigue	.43**	.37**	−.22, .10
								Ostracism	.40**	.59**	.01, .38
Managerial responsibility stress (.86)	.46**	150	13.89* (5)	.97	.03	**.47**	.22	**Emotional demands**	.41**	.57**	.02, .31
								Rumination (negative)	.41**	.39**	−.21, .17
								Quantitative workload	.33**	.27**	−.24, .11
Meaning (.94)	.74**	150	.04 (2)	1.00	.00	**.77**	.59	Intrinsic motivation	.55**	.53**	−.14, .11
								Demands-abilities job fit	.53**	.56**	−.08, .14
								Affective commitment	.53**	.56**	−.10, .17
Meeting effectiveness (.92)	.66**	144	6.90 (14)	1.00	.01	**.69**	.47	Innovation climate	.58**	.60**	−.11, .17
								Procedural justice	.51**	.45**	−.20, .07
								Supervisor competence	.48**	.52**	−.11, .20
Mental fatigue (.93)	.68**	145	32.18** (14)	.98	.03	**.71**	.50	Rumination (negative)	.47**	.59**	−.02, .26
								Loneliness	.41**	.34**	−.27, .12
								Time pressure	.29**	.60**	.17, .46
Needs-supplies job fit (.91)	.60**	118	11.04** (2)	.97	.03	**.62**	.38	Affective commitment	.59**	.67**	−.03, .18
								Perceived contract breach	−.50**	−.54**	−.21, .13
								Perceived leader effectiveness	.44**	.46**	−.18, .21
Negative effort-reward imbalance (.93)	.70**	122	19.96 (14)	.99	.02	**.74**	.55	Perceived contract breach	.53**	.65**	−.03, .26
								Organizational politics	.51**	.45**	−.22, .10
								Competitive goals	.48**	.50**	−.13, .17
Normative commitment (.81)	−.08	136	5.6 (2)	.97	.04	**−.18**	.03	**Organizational reputation**	−.33**	.33**	.46, .87
								Training climate	−.13	.39**	.30, .75
								Welfare climate	−.12	.46**	.36, .79
Organizational politics (.93)	.60**	137	25.39* (14)	.98	.03	**.64**	.40	Work frustration	.54**	.45**	−.24, .06
								Competitive goals	.52**	.75**	.10, .35
								Procedural justice	−.49**	−.48**	−.15, .17
Organizational reputation (.91)	.74**	120	3.64 (2)	1.00	.01	**.77**	.59	Welfare climate	.62**	.73**	−.02, .25
								Perceived contract breach	−.55**	−.64**	−.25, .06
								Person-organization fit	.48**	.54**	−.08, .21
Ostracism (.85)	.56**	129	21.78** (9)	.97	.03	**.60**	.36	Loneliness	.58**	.59**	−.15, .17
								Competitive goals	.39**	.41**	−.16, .18
								Abusive supervision	.36**	.19*	−.36, .02

Table 2 (continued)

Construct (MI alpha)	*r*	*N*	χ^2 (DF)	CFI	SRMR	Single-item		Criterion validity			
						Loading	Com	Construct	SI-MI	MI-MI	95% CI
Perceived contract breach (.87)	.75**	103	71.09** (9)	.89	.08	**.69**	.48	Welfare climate	−.60**	−.67**	−.20, .07
								Pessimism of org. change	.59**	.55**	−.18, .09
								Perceived org. support	−.51**	−.59**	−.23, .07
Perceived leader effectiveness (.94)	.77**	146	21.96** (5)	.98	.02	**.79**	.63	Leader warmth	.77**	.78**	−.07, .10
								Leader competence	.74**	.80**	−.02, .13
								Supervisor trust	.46**	.53**	−.05, .19
Perceived organizational support (.89)	.74**	120	4.13 (2)	.99	.02	**.78**	.61	Welfare climate	.73**	.74**	−.08, .09
								Perceived contract breach	−.68**	−.59**	−.05, .21
								Neg. effort-reward imbalance	−.61**	−.48**	.00, .27
Perceived overqualification (.68)	.50**	114	17.47** (5)	.92	.07	**.69**	.48	Neg. effort-reward imbalance	.33**	.35**	−.18, .21
								Daily hassles	.31**	.24*	−.29, .15
								Perceived contract breach	.21**	.38**	−.05, .39
Performance goal orientation (.71)	.18*	139	11.18 (9)	.98	.04	**.23**	.05	Cooperative orientation	.38**	.21*	−.38, .04
								Learning goal orientation	.32**	.24*	−.32, .16
								Deep acting	.31**	.18	−.37. .11
Person-organization fit (.87)	.73**	144	35.94** (5)	.93	.05	**.74**	.55	Needs-supply fit	.60**	.54**	−.22, .09
								Affective commitment	.58**	.62**	−.08, .15
								Meaning	.58**	.41**	−34, −.01
Perspective-taking (.83)	.39**	128	4.71 (5)	1.00	.02	**.42**	.18	Job self-efficacy	.35**	.21*	−.36, .07
								Quality of group experience	.34**	.22	−.37, .13
								Per. leader effectiveness	.32**	.36**	−.13, .21
Pessimism of organizational change (.89)	.62**	137	12.31* (5)	.98	.02	**.65**	.43	**Organizational reputation**	−.41**	−.60**	−.33, −.05
								Daily hassles	.40**	.36**	−.26, .13
								Organizational politics	.34**	.33**	−.19, .16
Physical fatigue (.95)	.69**	119	27.57* (14)	.98	.03	**.72**	.51	Subjective stress	.45**	.46**	−.17, .18
								Time pressure	.37**	.49**	−.05, .29
								Self-initiated work breaks	−.37**	−.50**	−.28, .02
Preference for group work (.91)	.81**	104	3.95 (2)	1.00	.01	**.83**	.69	Cooperative orientation	.56**	.53**	−.19, .12
								Cooperative goals	.23**	.26**	−.17, .22
								Meeting effectiveness	.20**	.42**	.10, .35
Procedural justice (.90)	.66**	133	52.98** (20)	.95	.05	**.69**	.47	Per. Leader effectiveness	.60**	.53**	−.20, .05
								Role conflict	−.54**	−.49**	−.11, .21
								Trust in supervisor	.41**	.51**	−.07, .26
Prosocial identity (.78)	.66**	107	5.73 (2)	.98	.03	**.76**	.58	Prosocial motivation	.53**	.41**	−.28, .05
								Perspective-taking	.37**	.15	−.42, −.01
								Cooperative orientation	.29**	.27**	−.21, .16

Table 2 (continued)

Construct (MI alpha)	*r*	*N*	χ^2 (DF)	CFI	SRMR	Single-item		Criterion validity			
						Loading	Com	Construct	SI-MI	MI-MI	95% CI
Prosocial motivation (.93)	.48**	116	14.87 (9)	.99	.03	**.48**	.23	Cooperative orientation	.53**	.48**	−.24, .15
								Perspective taking	.49**	.47**	−.17, .14
								Meaning	.49**	.25**	−.43, −.06
Quality of group experience (.93)	.78**	140	4.12 (2)	1.00	.01	**.79**	.63	**Climate for civility**	.61**	.80**	.09, .29
								Relationship conflict	−.61**	−.49**	−.04, .26
								Competitive goals	−.55**	−.71**	−.25. −.05
Quantitative workload (.83)	.74**	139	27.09** (9)	.96	.04	**.80**	.64	**Time pressure**	.61**	.83**	.13, .32
								Work pressure	.61**	.69**	−.04, .20
								Subjective stress	.46**	.56**	−.02, .23
Rational decision-making style (.88)	.45**	135	17.73* (9)	.97	.04	**.46**	.22	Perspective taking	.38**	.40**	−.24, .28
								Learning goal orientation	.37**	.42**	−.12, .23
								Cooperative goals	.32**	.41**	−.10, .28
Relationship conflict (.94)	.67**	114	21.97** (5)	.96	.03	**.66**	.44	Organizational politics	.59**	.54**	−.19, .08
								Daily hassles	.54**	.39**	−.38, .08
								Emotional demands	.50**	.35**	−.32, .03
Resources (.90)	.67**	152	23.34 (14)	.99	.03	**.70**	.49	*Unnecessary tasks*	−.54**	−.26**	.10, .45
								Innovation climate	.54**	.45**	−.26, .08
								Perceived org. support	.48**	.57**	−.05, .21
Role conflict (.84)	.56**	161	4.95 (5)	1.00	.02	**.60**	.36	Work frustration	.57**	.48**	−.25, .07
								Competitive goals	.55**	.66**	−.02, .25
								Emotional fatigue	.53**	.40**	−.29, .03
Rumination (negative) (.93)	.76**	121	10.32 (5)	.99	.02	**.78**	.60	Mental fatigue	.53**	.59**	−.10, .23
								Subjective stress	.49**	.66**	.05, .28
								Role conflict	.43**	.43**	−.17, .16
Self-initiated work breaks (.82)	.67**	133	.73 (2)	1.00	.01	**.74**	.54	Resources	.29**	.33**	−.12, .20
								Quantitative workload	−.25**	−.25**	−.18, .17
								Emotional demands	−.24**	−.53**	−.44, −.14
Subjective monotony (.84)	.69**	121	28.80** (5)	.93	.05	**.75**	.57	Meaning	−.57**	−.43**	−.01, .29
								Needs-supply fit	−.44**	−.35**	−.09, .26
								Cognitive demands	−.41**	−.32**	−.09, .25
Subjective stress (.83)	.79**	140	15.89** (5)	.98	.04	**.85**	.73	Work frustration	.58**	.63**	−.07, .17
								Quantitate workload	.57**	.56**	−.12, .10
								Time pressure	.55**	.61**	−.07, .18
Supervisor competence (.94)	.79**	146	55.64** (14)	.96	.03	**.81**	.66	Informational justice	.74**	.67**	−.18, .04
								Perceived contract breach	−.52**	−.50**	−.12, .17
								Trust in supervisor	.48**	.36**	−.29, .06
Supervisor warmth (.95)	.86**	139	52.19** (14)	.97	.02	**.87**	.76	Perceived leader effectiveness	.79**	.78**	−.09, .07
								Interpersonal justice	.75**	.80**	−.02, .12
								Abusive supervision	−.54**	−.54**	−.15, .15

Table 2 (continued)

Construct (MI alpha)	*r*	*N*	χ^2 (DF)	CFI	SRMR	Single-item		Criterion validity			
						Loading	Com	Construct	SI-MI	MI-MI	95% CI
Surface acting (.85)	.69**	131	3.59 (2)	.99	.02	**.77**	.59	*Daily hassles*	.53**	.35**	− .33, − .03
								Organizational politics	.50**	.36**	− .33, .05
								Competitive goals	.45**	.46**	− .15, .18
Task conflict (.87)	.59**	139	17.59** (5)	.96	.03	**.63**	.39	Daily hassles	.54**	.39**	− .30, .00
								Competitive goals	.48**	.44**	− .23, .15
								Relation conflict	.44**	.58**	− .02, .29
Task interdependence (.70)	.17	122	4.14 (2)	.97	.04	− .20	.04	Cooperative orientation	.37**	.19	− .45, .10
								Cooperative goals	.35**	.11	− .51, .02
								Relation conflict	− .03	− .15	− .34, .10
Team self-management (.84)	.54**	100	.66 (2)	1.00	.01	**.57**	.32	Autonomy climate	.43**	.46**	− .18, .24
								Coworker trust	.29**	.36**	− .11, .23
								Quality of group experience	.21**	.30**	− .11, .30
Team workload sharing (.80)	.56**	125	4.09 (2)	.99	.03	**.65**	.43	Meeting effectiveness	.51**	.43**	− .26, .12
								Quality of group experience	.46**	.57**	− .03, .26
								Coworker trust	.45**	.48**	− .15, .20
Time pressure (.82)	.53**	131	6.82 (5)	.99	.03	**.59**	.34	**Quantitative workload**	.53**	.83**	.20, .41
								Work pressure	.47**	.69**	.09, .37
								Subjective stress	.43**	.61**	.04, .32
Training climate (.95)	.74**	100	27.84* (14)	.98	.02	**.75**	.56	Innovation climate	.60**	.73**	.00, .27
								Resources	.54**	.44**	− .26, .07
								Perceived org. support	.53**	.62**	− .06, .24
Trust in supervisor (.89)	.52**	123	3.99 (2)	.99	.02	**.56**	.31	*Per. leader effectiveness*	.80**	.53**	− .38, − .16
								Supervisor competence	.68**	.36**	− .48, − .15
								Informational justice	.66**	.48**	− .32, − .04
Unnecessary illegitimate tasks (.81)	.58**	132	8.51 (5)	.99	.03	**.65**	.42	Role conflict	.64**	.54**	− .23, .03
								Neg. effort-reward imbalance	.60**	.55**	− .19, .08
								Efficiency climate	− .52**	− .42**	− .06, .26
Unreasonable illegitimate tasks (.89)	.49**	118	4.82 (5)	1.00	.02	**.52**	.27	Role conflict	.55**	.56**	− .12, .16
								Innovative climate	− .51**	− .28**	.05, .42
								Neg. effort-reward imbalance	.49**	.61**	− .04, .27
Welfare climate (.87)	.79**	106	7.26 (5)	.99	.02	**.81**	.65	Perceived contract breach	− .74	− .67**	− .05, .20
								Person-organization fit	.64**	.71**	− .03, .18
								Quality of group experience	.61**	.68**	− .03, .17
Work authenticity (.84)	.64**	116	30.31** (2)	.90	.06	**.63**	.40	Meaning	.46**	.36**	− .27, .08
								Intrinsic motivation	.42**	.28**	− .35, .09
								Career satisfaction	.32**	.57**	.08, .42

Table 2 (continued)

Construct (MI alpha)	*r*	*N*	χ^2 (DF)	CFI	SRMR	Single-item		Criterion validity			
						Loading	Com	Construct	SI-MI	MI-MI	95% CI
Work frustration (.74)	.66**	141	1.77 (2)	1.00	.01	**.73**	.53	Neg. effort-reward imbalance	.60**	.55**	−.18, .08
								Subjective stress	.59**	.63**	−.09, .16
								Rumination (negative)	.54**	.46**	−.22, .07
Work hypercompetitive (.81)	.24*	107	1.17 (2)	1.00	.02	**.26**	.07	Competitive orientation	.42**	.45**	−.17, .23
								Task conflict	.22**	.35**	−.11, .36
								Performance goal orientation	.20*	.49**	.07, .50
Work pressure (.81)	.66**	130	4.74 (2)	.99	.02	**.71**	.50	Quantitative workload	.67**	.69**	−.09, .13
								Time pressure	.61**	.69**	−.05, .21
								Emotional demands	.57**	.37**	−.35, −.04

r: correlation between single- and multi-item measures. N: CFA sample size. DF: degrees of freedom, CFI: Central fit index. SRMR: Standardized root mean residual. Bold loading: significant at least $p < .05$. Com.: Communality. SI-MI: correlation between the single-item measure of the focal construct and the multi-item measure of the criterion construct. MI-MI: correlation between the multi-item measure of the focal construct and the multi-item measure of the criterion construct. 95% CI: confidence interval around the correlation difference. Bolded criterion construct: 95% does not include zero in favor of the multi-item measure. Italicized criterion construct: 95% does not include zero in favor of the single-item measure

* $p < .05$; ** $< .01$

the convergent validity estimate for *the single-item measure* of the focal construct with the multi-item measure of the criterion construct and the convergent validity estimate for *the multi-item measure* of the focal construct with the multi-item measure of the criterion construct. Put another way, while we are interested in potential differences in criterion validity across the two assessment methods, we leverage the "convergent validity" estimate within the MTMM approach to empirical test for criterion validity. Per Raykov, we used the Delta method within Mplus. The resulting 95% confidence interval for the correlational difference is indicative of the degree to which the two convergent validity estimates (i.e., the convergent validity of the single- or multi-item measure of the focal construct for the criterion variable) are meaningfully different. That is, if the 95% confidence interval for the correlational difference (see Table 2) includes zero, the two convergent validity estimates (i.e., the convergent validity of the single- or multi-item measure of the focal construct for the criterion variable) do not meaningfully differ from one-another; the two measurement approaches demonstrate approximately equal criterion validity. Again then, while we report on differences in convergent validity estimates within a MTMM matrix, we are in effect using these estimates to examine differences in criterion validity across the two measurement approaches.

Per Table 2, to provide a robust examination of our Research Question, we examined three criterion constructs for each single-item measure. While beyond our scope for a detailed discussion, the criterion constructs were selected based on existing conceptual and empirical work specific to a given focal construct; our ability to match to theoretically relevant criterion constructs was meaningfully limited by the constructs included in the overall data collection effort and should receive additional attention in future research on the applicability of single-item measures within the organizational sciences.

For simplicity, each focal-criterion construct pairing was examined in its own model (Raykov 2011). In these analyses, we sought to leverage all available data to produce more accurate standard errors (Newman 2014). Thus, if a respondent completed at least one of the four possible measures (i.e., either the single- or multi-item measures of the focal construct, and/or the single- or multi-item measures of the criterion construct), they were retained. Full information maximum likelihood estimation was used to account for missing data (Newman 2014); each MTMM matrix was estimated based on a sample of 1000 or more respondents.

In terms of our Research Question, results from the MTMM matrix analyses, and the resulting Delta 95% confidence intervals, are summarized in Table 2. Collectively, 273 MTMM convergent validity estimates were examined (three for each single-item measure). In approximately 12% of these analyses, the convergent validity estimate for the

multi-item measure of the focal construct was more strongly correlated to the criterion measure (Table 2) than the single-item measure of the focal construct. In approximately 7% of these analyses, the convergent validity estimate for the single-item measure of the focal construct was more strongly correlated to the criterion measure (denoted in italics in Table 2) than the multi-item measure. For the remaining 81.3%, there was no meaningful difference in the convergent validity coefficients, suggesting that, yes, the single-item measures demonstrate criterion validity that is generally comparable to multi-item measures of the same construct.

That said, per Smith et al. (2000) in their discussion of short-form measures, another way to conceptualize our Research Question is in terms of choosing "the best balance between time-resource savings and loss of validity" (p. 110). That is, scholars may be willing to "trade" (i.e., accept) some reduction in, for example, criterion validity, in return for a short measure that (a) places less burden on respondents (assuming the shorter measure is still reliable and content valid; Smith et al. 2000) while (b) still defensibly measuring the construct of interest[9] (Cortina et al. 2020). Thus, we examined the omnibus reduction in criterion validity. That is, while 81.3% of the 273 relationships examined above were not statistically different, there could still be a meaningful downward bias (this in addition to the 12% of relationships where criterion validity was statically stronger for the multi-item measures). Collectively then, as a measurement approach, there could be systematic loss in criterion validity when single-item measures are used over multi-item measures of the same construct. To examine this issue, we computed the average difference between the two convergent validity estimates (i.e., for the single- vs multi-item measure) across the 273 comparisons reported in Table 2. On average, the observed convergent validity estimates for single-item measures was 0.02 ($SD = 0.15$) lower than for convergent validity estimates for multi-item measures. That is, while there is considerable variability for given constructs, in using a single-item measure, on average, researchers are "trading away" 0.02 in criterion validity (at least based on the data examined in study 4).

Collectively then, it is important to recognize that the observed criterion validity results are two-sided. On the one hand, as a general approach, using single-item measures may not result in "trading away" criterion validity. However, when zeroing in on a given construct, switching from a multi-item measure to a single-item approach might result in *significant* reduction in criterion validity. As such, again, we stress that our results should not be interpreted as suggesting all construct can or should be measured with a single-item measure — such decisions must be made on a construct-by-construct basis.

Supplementary Analyses

There is a standing tradition of estimating a "consistency" reliability for single-item measures based on item communalities, wherein scholars square the standardized CFA factoring loading of single-item measures (i.e., Hypothesis 4; see Spörrle and Bekk 2014). In Table 2, we report the consistency-based reliability for each single-item measure based on this "communality" approach (Wanous and Reichers 1996; Wanous et al. 1997). However, it is important to recognize that this estimate is dependent on the multi-item measure used (Spörrle and Bekk 2014); if the multi-item measure used is deficient in some way (either conceptually or psychometrically), it will influence the estimated consistency-based reliability of the single-item measure. For example, Spörrle and Bekk demonstrated that the more items in the multi-item reference measure, the lower the single-item measure consistency-based reliability. To this end, we conducted a series of analyses based on the content validity evidence from study 1 and the psychometric information for both measurement approaches in study 4.

Content validity estimates for the multi-item measures positively correlated with estimates of internal consistency ($r = 0.34$, $p < 0.001$); the more content valid raters perceived the items in the multi-item measure, the more likely people were to respond to those items in an internally "consistent" way. In turn, multi-item measure internal consistency estimates correlated at 0.39 ($p < 0.001$) with the single-item consistency-based reliability estimates. With that in mind, we applied Hayes' (2018) PROCESS procedure and determined that multi-item content validity had an unstandardized indirect effect of 0.29 (95% CI: 0.07, 0.56) on single-item consistency-based reliability by way of the internal consistency estimate for multi-item measures. The implication is that applying multi-item measures with weaker psychometric characteristics can result in an underestimation of consistency-based reliability estimates for single-item measures.

Study 5

Study 4 results suggest that, generally speaking, many of the proposed single-item measures demonstrate acceptable construct validity, such that they "load" with items from multi-item measures of the same construct. In turn, the single-item measures demonstrated criterion validity that was comparable to multi-item measures of the same construct, with

[9] Relatedly, in an effort to reduce the trade-off between the number of items in the scale length and the scale quality, Cortina et al. (2020) have developed a procedure that aims to optimize the scale quality (e.g., alpha reliability coefficient, part-whole correlations) of the resulting shortened scale by analyzing all possible sets of items drawn from the full scale.

minimal downward bias (loss) in criterion validity (Smith et al. 2000). Combined with results from studies 1–3, in Table 3, based on our collective evidence, we provide an overarching evaluation of the triangulated reliability and validity evidence for the single-items measures.[10]

Four measures demonstrated no validity in that these measures had limited to low content validity, some usability concerns, lower test–retest reliability, and limited criterion validity. On the other hand, 56 measures demonstrated *very good* validity in that they evidenced moderate to high content validity, no usability concerns, moderate to high test–retest reliability, and extensive criterion validity. Beyond that, another 19 measures demonstrated *extensive* validity; these measures evidenced high content validity, no usability concerns, systematically high test–retest reliability, and extensive criterion validity. With this evidence in mind, we now seek to pivot and take a forward-looking perspective on the application of single-item measures.

As noted, there is a host of work discussing under what conditions the use of single-item measures is appropriate. Perhaps the most influential treatment of this issue, in part because of the checklist provided, is the work by Fuchs and Diamantopoulos (2009). A core premise put forth by Fuchs and Diamantopoulos and others (e.g., Rossiter 2002, 2008) is that as construct breadth increases, single-item measures become less reliable and valid. This is because single-item measures require respondents to assess the overall construct, and as the construct becomes conceptually broad and complex, they are likely to interpret the construct ambiguously and to ignore essential aspects of the construct when answering the single-item measure (Fuchs and Diamantopoulos 2009). In this case, multi-item measures help respondents assess essential aspects of a construct and researchers can subsequently combine item-level responses to capture their overall standing on the underlying construct (Fuchs and Diamantopoulos 2009).

[10] We established construct validity evaluations based on a point system; we assigned points based on content validity (i.e., 1 point: content validity ≤ .69, 5 points: content validity ≥ .90), amount of usability concerns (i.e., 1 point: systematic usability concerns, 4 points: no meaningful usability concerns), average ICC test–retest reliability scores (i.e., 1 point: ICC < .40, 4 points: ICC > .74), construct validity (i.e., 0 points: CFA factor loading < .32, 5 points: CFA factor loadings > .70), and criterion validity (i.e., 1 point: limited to no evidence of criterion validity, 5 points: systematic evidence of criterion validity). We then computed an average across these different pieces of reliability and validity (scores ranged from 1.67 to 4.47). Constructs with scores greater than 4.00 were evaluated as having *extensive* construct validity, constructs between 3.00 and 3.99 were evaluated as demonstrating *very good* construct validity, constructs between 2.70 and 2.99 were evaluated as demonstrating *good* construct validity, constructs between 2.25 and 2.69 were evaluated as demonstrating *limited* construct validity, and constructs less than 2.25 were evaluated as demonstrating *no* construct validity. Additional information is available upon request.

Constructs can be evaluated as existing along a *construct complexity* continuum (Fuchs and Diamantopoulos 2009; Rossiter 2002), where at one end, constructs are conceptually narrow (i.e., simple, unidimensional). At the other end are constructs that are more complex, or conceptually broad (e.g., multi-dimensional). As constructs become simpler (i.e., narrowly defined), the general argument is that fewer items are needed to adequately represent the conceptual space of a given construct (Allen et al. 2022). On the other hand, as constructs become more complex or conceptually broad, more distinct content (and items representing that content) is needed to appropriately represent each aspect of the construct (Fuchs and Diamantopoulos 2009). For purposes of contextualization then, it is here where much of the resistance to single-item measurers originates in the journal review process, wherein reviewers and journal editors will indicate the perspective that a given construct is *too broad* or *complex* to be assessed with a single-item measure based on the conceptual definition of the construct.

Implicitly then, if scholars develop a thorough understanding of a given construct's conceptual breadth, they will be well positioned to know if that construct can be assessed with a single-item measure. As noted previously though, to our knowledge, there is no established procedure for estimating the degree to which a construct is "broad" versus "narrow." Again, scale authors are generally encouraged to rely on their "professional judgement" (Gehlbach and Brinkworth 2011, p. 383) when deciding on the number of items needed to effectively represent a construct, relative to the conceptual breadth of the target construct (Hinkin 1998). For this reason, rather than using potential construct breadth as a criterion for deciding which construct to include in the current program of research, we take an approach to empirically assess the level of conceptual breadth of constructs and examine how it relates to the reliability and validity of single-item measures. Specifically, in study 5, we leverage subject matter expert evaluations of construct breath and seek to empirically examine the potential boundary conditions that construct complexity places on the application of single-item measures. Based on the established arguments (e.g., Fuchs and Diamantopoulos 2009), we predict the following:

Hypothesis 5 *Construct breadth (indexed based on subject matter expert ratings) is negatively related to reliability and validity evidence for single-item measures.*

Method

Participants and Procedure

A total of 103 subject matter experts (SMEs) were recruited directly by the study authors, as well as by placing a request to participate via a research listserv (respondents were also

Table 3 Triangulated single-item validity evaluations with SME construct breadth rating (study 5) reported in parenthesis

Extensive validity		
Abusive supervision (2.58)	Affective commitment (2.10)	Coworker trust (2.54)
Emotional fatigue (2.18)	Innovation climate (2.94)	Interpersonal justice (2.46)
Loneliness (2.60)	Meaning (3.83)	Mental fatigue (2.62)
Perceived leader effectiveness (3.70)	Perceived organizational support (3.50)	Preference for group work (2.47)
Prosocial identity (2.54)	Quantitative workload (2.14)	Subjective monotony (1.89)
Supervisor competence (2.95)	Supervisor warmth (2.50)	Unnecessary illegitimate tasks (2.41)
	Welfare climate (3.79)	
Very good validity		
Authoritarian leadership (2.36)	Autonomy climate (2.80)	Bureaucracy (3.42)
Career satisfaction (3.08)	Climate for civility (3.00)	Cognitive demands (3.04)
Competitive goals (3.11)	Cont. commitment-low alternatives (1.39)	Cooperative goals (3.07)
Cooperative orientation (2.50)	Daily work hassles (3.21)	Demands-abilities job fit (3.26)
Distributive justice (2.42)	Emotional demands (2.67)	Family authenticity (2.77)
Goal-focused leadership (2.15)	Informational justice (1.97)	Intrinsic motivation (2.46)
Intuitive decision-making style (2.65)	Job Insecurity (2.00)	Job self-efficacy (2.46)
Leader avoidant conflict behaviors (2.25)	Leader collab. conflict behaviors (2.55)	Leadership self-identity (2.47)
Learning goal orientation (2.72)	Meeting effectiveness (2.71)	Needs-supplies job fit (3.40)
Negative effort-reward imbalance (2.75)	Organizational politics (4.00)	Organizational reputation (3.80)
Ostracism (2.33)	Perceived contract breach (2.74)	Perceived overqualification (2.40)
Person-organization fit (3.33)	Physical fatigue (2.44)	Procedural justice (2.07)
Prosocial motivation (2.54)	Quality of group experience (3.31)	Rational decision-making style (2.63)
Relationship conflict (3.18)	Resources (4.38)	Role conflict (2.79)
Rumination (negative) (2.04)	Self-initiated work breaks (1.76)	Subjective stress (3.75)
Surface acting (2.12)	Task conflict (2.54)	Team self-management (2.65)
Team workload sharing (2.94)	Time pressure (1.96)	Training climate (3.53)
Trust in supervisor (2.81)	Unreasonable illegitimate tasks (2.64)	Work authenticity (2.90)
Work frustration (2.22)	Work pressure (2.73)	
Good validity		
Formalization climate (4.00)	Managerial responsibility stress (2.92)	Perspective-taking (2.52)
Limited validity		
Competitive orientation (2.30)	Cont. commitment — high sacrifices (1.55)	Efficiency climate (3.27)
Extrinsic motivation (3.07)	Family motivation (2.62)	Leader domin. conflict behaviors (1.80)
Performance goal orientation (2.54)	Pessimism of organizational change (2.64)	Work hypercompetitive (2.17)
No validity		
Deep acting (2.14)	Face-time orientation (3.00)	Normative commitment (2.04)
	Task interdependence (2.21)	

Extensive validity: high content validity, no usability concerns, systematically high test–retest-reliability, extensive criterion validity. *Very good validity*: moderate to high content validity, no usability concerns, moderate to high test–retest-reliability, extensive criterion validity. *Good validity*: moderate content validity, limited usability concerns, moderate test–retest-reliability, systematic criterion validity. *Limited validity*: low to moderate content validity, limited usability concerns, low to moderate test–retest-reliability, systematic albeit weaker criterion validity. *No validity*: limited to low content validity, some usability concerns, low test–retest-reliability, poor criterion validity

asked to forward the request to participate to other researchers). Approximately two-thirds of the SMEs (65.1%) held a Ph.D. or other professional degree (e.g., JD, MD); the remaining third were graduate students. Approximately three-fourths of the sample (75.6%) reported that their academic training was in industrial/organizational psychology. On average, SMEs had published 21.7 ($SD = 28.12$) academic manuscripts with half of all SMEs having previously been involved in publishing scale development and/or scale validation work.

SMEs were first presented with background information around the concept of construct breadth. Specifically, SMEs were informed that construct breadth exists along a continuum and that "'Broad' constructs, based on their conceptual definition, have a large content space. Broad constructs are sometimes referred to as complex, abstract, fuzzy, or

large bandwidth constructs." In turn, they were informed that, "'Narrow' constructs, again based on their conceptual definition, have a smaller content space. Narrow constructs are sometimes referred to as concrete, focused, simple, or narrow bandwidth constructs." Based on the constructs' definitions from study 1, SMEs were then randomly presented with 30 constructs to evaluate.

Measures

After reading a randomly assigned construct definition, SMEs were asked to make three judgments: *construct familiarity*, *definitional adequacy*, and *construct breadth*. To measure construct familiarity, they were asked, "How familiar are you with this construct?" Responses were made on a 5-point scale (i.e., 1 = *Not at all*, 5 = *Extremely*). To measure definitional adequacy, they were asked, "Does this conceptual definition adequately represent the construct?" Responses were made on a 5-point scale (i.e., 1 = *Poor*, 5 = *Excellent*). To measure construct breadth, SMEs were asked, "To what degree do you think this is a conceptually narrow construct or a conceptually broad construct?" Responses were made on a 5-point scale (i.e., 1 = *Very narrow*, 5 = *Very broad*).

Results

To better ensure accuracy of SME judgments regarding the adequacy of a given construct definition, and evaluations of construct breadth, we only retained respondents who indicated they were at least *somewhat* familiar (i.e., 3.00 or above) with the given construct for analysis purposes; on average, each construct was evaluated by 20.78 ($SD = 6.58$) SMEs.

All constructs were evaluated by SMEs as having at least good definitional adequacy (3.00 or above); the average definitional adequacy rating was 3.90 ($SD = 0.28$) with complete details reported in the Online Supplemental Materials. Table 3 includes SME construct breath ratings, wherein ratings ranged from 1.39 for *Continuance commitment-low alternatives* to 4.40 for *Resources*; the average construct breadth rating was 2.71 ($SD = 0.58$).

To test Hypothesis 5, we ran a series of correlations between SME construct breadth ratings and reliability and validity information collected from studies 1–4. SME construct breadth ratings were unrelated to single-item content validity ratings from study 1 ($r = 0.00$, $p = 0.991$) and were unrelated to the summated usability issues score from study 2 ($r = 0.03$, $p = 0.746$). SME construct breadth ratings were also unrelated ICC test–retest reliabilities from study 3 (condition 1, $r = 0.00$, $p = 0.974$; condition 2, $r = 0.16$, $p = 0.125$; condition 3, $r = 0.10$, $p = 0.358$). From study 4, there was no relationship between SME construct breadth ratings and single-item construct validity estimates based on the observed CFA factor loading ($r = 0.09$, $p = 0.370$). There was also no relationship ($r = 0.13$, $p = 0.238$) between SME construct breadth ratings and convergent validity estimates (i.e., the degree to which scores across the two measurement approaches correlated for the same given construct).

Finally, we examined if broader constructs, based on SME ratings, resulted in a greater downward bias in criterion validity scores. As a reminder, per study 4 (and reported in Table 2), three criterion estimates were generated for each construct. Within each construct, for each comparison, the multi-item criterion validity estimate was subtracted from the single-item criterion validity estimates. In turn, the three comparisons were averaged to create an overall score. For example, across the three criterion comparisons for *abusive supervision*, there was an overall downward bias of 0.03. Across the 91 constructs, there was no relationship between differences in overall criterion validity and construct breadth ($r = 0.20$, $p = 0.058$).[11]

There is an important issue that should be considered though when interpreting these results. While SME construct breath ratings ranged from 1.39 to 4.40, as we noted when describing our construct selection process, we intentionally disqualified constructs that had inconsistent conceptual definitions in the literature when selecting constructs for the current research. Doing so may have resulted in construct breadth range restriction in that there may be more inconsistency in defining "broader" constructs (e.g., Casper et al. 2018). This range restriction may have influenced the nature of our results and may serve as an important area for future research.

General Discussion

Through this evidence-based program of research, we demonstrate that, yes, for many constructs, single-item measures are a reliable and valid measurement approach. Put another way, the majority of the single-item measures under consideration here were both reliable and valid measures of the underlying construct based on a systematic triangulation methodology. It is clear then, moving forward, that it is incumbent on researchers and reviewers alike to evaluate the application of single-item measures in a given context and not rely on subjective biases about their generalized reliability and validity. The use of single-item measures *is not* an inherent indicator of a weak research design, nor are researchers inherently trading away validity for convenience. To be clear, our results do not support the argument that all

[11] There was also no relationship between SME ratings of construct breadth and the final triangulated construct validity level reported in Table 3 ($r = 0.15$, $p = 0.14$).

constructs can be measured with single-item measures, nor that all single-item measures are valid. As research in the organizational sciences expands and evolves, the application of single-item measures within a given study should be evaluated relative to the challenges their use might help address. Admittedly, this is not a new recommendation given that a list of conditions under which single-item measures may work appropriately has been offered in the literature (e.g., Fuchs and Diamantopoulos 2009). However, our research further contributes to the literature by providing the most comprehensive evidence-based review of the issue to date (Allen et al. 2022).

More practically, we provide a compendium of single-item measures that researchers can draw on *without* sacrificing their ability to validly assess the relevant constructs. Collectively, these measures seem particularly well-suited to help scholars address a myriad of conceptual, methodological, and empirical challenges within their research. Below we highlight some of our primary findings with an eye towards what our findings mean in terms of the application of single-item measures. In turn, we outline a general process for researchers to leverage when developing and validating single-item measures. We then consider limitations and additional future research directions related to our program of research.

Primary Findings

Perhaps one of the most surprising findings was how few of the single-item measures demonstrated usability concerns. Given single-item measures may be more cognitively demanding (given their length, and restatement of construct definitions), we had anticipated observing systematic usability concerns for at least some items, wherein the reading level of a given item may help explain those possible usability concerns; however, systematic usability concerns did not manifest.

Although classic wisdom suggests simpler, less cognitively demanding items are better, empirical work in educational and employment testing on subgroup differences as a function of the readability of test items offers a more nuanced story (Freedle 2003; Freedle and Kostin 1992, 1997; Scherbaum and Goldstein 2008). It is interesting then to consider our finding that items with higher reading levels demonstrated lower test–retest reliability across the three conditions. While beyond the scope of the current research, it may be that these more complicated items (Fowler 1995) resulted in greater specificity (i.e., reduced ambiguity) and increased accuracy in assessing change over time resulting in lower test–retest reliabilities. To this end, reading level and content validity estimates were positively correlated ($r = 0.24$, $p = 0.02$) potentially suggesting that more complicated items are more accurately tapping into the construct (Guion 1965).

Abstracting these findings, we would argue that the construction of single-item measures is an artful balancing act of two inter-related factors. First, it is necessary to construct single-item measures that reduce ambiguity to ensure that the items are consistently understood, while still being content valid. Beyond that, items must be written to ensure all participants have access to the information needed to answer the question accurately. These efforts may inherently mean increasing readability scores by providing additional, necessary detail (through more words and/or more complex sentence structures). While there were no meaningful differences in usability concerns based on reading level, we would encourage researchers to continue to consider if issues like age and language proficiency (Tourangeau et al. 2020) influence the utility of single-item measures across different populations.

Returning to a previous point, while certainly efforts can be undertaken to refine and reduce the cognitive demand of single-item indicators (e.g., reduce word length, reading level) in hopes of optimizing reliability, we are left questioning whether there is potential risk in over-indexing on test–retest reliability as an indicator of the validity and utility of a single-item measure. Consider the single-item measure of supervisor interpersonal justice, "My supervisor was generally respectful and polite when discussing work related issues with me." Despite the item's *very strong* definitional correspondence relative to its multi-item counterpart (0.91 and 0.83, respectively), this item's relatively high reading level (13.4) and marginally acceptable test–retest reliability (0.59, 0.71, 0.41, across conditions, respectively) may be deemed problematic by some. However, these reliability indices may be interpreted as advantageous, indicating the item's sensitivity to detect true score (episodic) change in perceptions of supervisors' interpersonal justice behaviors over time. To this end, it is important to recognize that numerous factors have been shown to influence the reliability of survey responses, including respondent characteristics (e.g., education, conscientiousness, household income), item-level characteristics (e.g., social desirability concerns), and temporal lag (e.g., Tourangeau et al. 2020). More research on the interaction of these various issues in the context of single-item measurement appears warranted.

Turning to our criterion-related validity evidence, the unified validity model, wherein content, construct, and criterion validity are inseparable (Messick 1995), is contested and in turn, scholars have advanced reconfigured guidelines for validating psychometric measures (see Hughes 2018). For example, in Hughes' (2018) two-step model, psychometric developers are charged with answering two fundamental psychometric questions sequentially: "am I measuring what I want to measure?", as supported by content, response

process, and structural evidence, and "is my measure useful?" as supported by convergent, discriminant, and concurrent validity evidence, among other types of evidence (Hughes 2018, p. 752). Guided, in part, by this approach to validation, we assessed whether our measures met or exceeded the criterion-related validity of their multi-item measure only once sufficient validity evidence was accumulated regarding the single items' accuracy and reliability in measuring their intended construct.

While programmatic and apparent from the logic of this paper, explicitly acknowledging this decision sets an important backdrop for interpreting the MTMM matrix findings reported in study 4, whereby approximately 88% of the criterion-related validity estimates for the single-item measure were comparable or exceeded criterion-related validity estimates of their multi-item measure counterparts. The robust criterion-related evidence for the single-item measures provides a compelling counterargument against those who say that the ceiling of single-item measures' criterion-related validity is inferior to their multi-item measures counterparts due to lower (a) reliability (Ziegler et al. 2014) and (b) content adequacy (Cheah et al. 2018). Instead, our findings suggest that intentional and rigorous development, refinement, and psychometric testing of single-item measures in ways that maximize reliability and content adequacy can yield criterion validity estimates for single-item measures that are comparable to multi-item measures.

Finally, specific to study 5, it is genuinely surprising that there appears to be no systematic relationship between SME evaluations of construct breadth and reliability and validity evidence, at least based on the 91 constructs examined here and the empirical evidence from studies 1 through 4. This result is inconsistent with researchers' (e.g., Fuchs and Diamantopoulos 2009) suggestions that single-item measures may be inappropriate to capture constructs that are conceptually broad, fuzzy, and complex because single-item measures of these constructs may be ambiguously interpreted without consideration of each of the essential facets of the constructs.

We see these (lack of) findings as a double-edge sword. On the one hand, it seemingly reinforces the argument that single-item measures are more applicable than is commonly accepted in the literature to date. That is, if researchers can provide sufficient reliability and validity evidence for a single-item measure for a given construct, that evidence should be evaluated, seemingly, irrespective of the over conceptual complexity of the construct. However, again, researchers are advised against arguing that just because a construct is "conceptually narrow," it is possible to develop a single-item measure of that construct — reliability and validity evidence must still be collected and presented, even for conceptually narrow constructs.

Admittedly, independent of whether the construct is conceptually broad and/or narrow, there is risk (by virtue of item content/phrasing) that the holistic interpretation of its single-item measure differs from that of its multi-item counterpart, but our results suggest this risk is not more prevalent with broad constructs (as may be expected with more broad constructs having more conceptually divergent item content across the multi-item measure set). However, there may exist, beyond SME evaluations, other methods to index construct breadth that may result in a more nuanced understanding of the potential boundary conditions it creates. We see this as a potentially fruitful area of research, both in terms of not only single-, but also, multi-item measures given the lack of concrete advice around the number of items needed to "validly" assess a construct.

Recommended Process to Validating Single-Item Measures Based on Lessons

An additional contribution to the literature from our program of research is the provision of a template others may leverage when (a) validating new single-item measures as well as (b) accumulating additional structural and external validity evidence to support inferences drawn from the single-item measures herein (Flake 2021; Flake and Fried 2020). As noted by Aguinis et al., (2021), "many published articles in management and related fields do not include sufficient information on precise steps, decisions, and judgment calls made during a scientific study" (p. 679). Thus, in the interest of transparency, we outline general steps others may wish to consider, informed by lessons learned while conducting this program of research.

First, while seemingly obvious, we encourage care to be taken to understand the conceptual definition of a construct. What became abundantly clear in our initial work was that there are many subtle and sometimes glaring differences in how many constructs are defined; consider, over 100 conceptual definitions exist for work-family balance (Casper et al. 2018), wherein different definitional approaches can lead to meaningfully different results (Wayne et al 2022). While subtle differences in defining a construct can be accounted for, without a firm understanding of the conceptual construct, it is difficult to develop a valid measure of that construct. Thus, while a construct like work-family balance may be on the extreme end, it was not the only construct excluded for poor conceptual clarity. We echo the call by others (e.g., Stone-Romero 1994) to avoid wasting time and effort by ensuring that the conceptual underpinnings of a construct are understood before engaging in scale development work.

From this conceptual understanding, it is time to write an item. Needless to say, there are a host of referred works that provide excellent recommendations for writing items

(e.g., Fowler and Cosenza 2009; Haladyna and Rodriguez 2013; Robinson and Leonard 2018) as well as ensuring those items are evaluated for conceptual disconnects (e.g., Hughes 2018). With those resources in mind, specific to writing single-item measures, we would encourage researchers to proactively consider both the response scale and associated recall window. For our purposes, we selected either a frequency or agree-disagree format. A particularly salient issue during the item writing process was the number of constructs that could be assessed with either response scale, with only minor wording revisions. We relied on the construct definition to provide guidance, conceptually, in terms of the appropriate response scale. However, as per ongoing discussions around addressing issues of common method variance (i.e., Spector et al. 2017), we would encourage researchers to choose a response scale (e.g., frequency, extent) that is considered to best fit with the nature of each item and to allow for multiple different response scales across items, as a potential solution to common method variance.

Once developed, it is necessary to implement a validation strategy, a process as much art as science. We encourage scholars to demonstrate the validity of their single-item measure through triangulation, focusing on types of validity that make the most sense relative to the construct(s) under consideration. We emphasized *definitional correspondence*, however, per Colquitt et al. (2019), depending on the construct(s) of interest, *definitional distinctiveness*, or rather, "the degree to which a scale's items correspond more to the construct's definition than to the definitions of other orbiting constructs" may be appropriate (p. 1243). Regardless, demonstrating that single-item measures are content valid is important (Schriesheim et al. 1993). In retrospect though, building from study 1, if the goal is to demonstrate that a single-item measure has a similar level of content validity as a multi-item measure of the focal construct, it may be more appropriate to have respondents evaluate the definitional correspondence of the items from the multi-item measures as a set. That is, taking a gestalt perspective, raters would evaluate the degree to which the set of items tap into, or represent, the focal construct's conceptual definition. In turn, if it can be shown that a single-item measure has similar definitional correspondence by asking just one item versus, for example, asking six, that would support the argument for content validity of the single-item measure.

In light of our goal of providing a robust understanding of single-item measures, we reported on the test–retest reliability of the proposed measures over three temporal conditions. While examining test–retest reliability is more resource intensive than estimating, for example, consistency reliabilities (Spörrle and Bekk 2014), that approach likely underestimates the reliability of the single-item measures. Furthermore, as we demonstrated in study 4, consistency-based reliabilities are inherently linked to the psychometric characteristics of multi-item reference measures. And more practically, given the nature of the construct, a valid multi-item measure of the focal construct may not exist. As such, while demonstrating reliability is a piece of the puzzle, and must be examined, it is difficult to provide specific guidance on which type of reliability might be most appropriate for a given construct or program of research. However, we would encourage researchers, and reviewers alike, to recognize that reliability is only a piece of the puzzle. That is, "lower" consistency-based reliabilities may be a reflection of weaknesses in the referent multi-item measure (see study 4), whereas "lower" test–retest reliabilities may be the expected and desired outcome for temporal sensitive constructs (e.g., mood) examined over various temporal lags in that said items are detecting true score change over time.

Continuing with the notion of triangulation, the heterogeneity of constructs included here allowed for an ideal examination of criterion validity. We were particularly pleased with the application of MTMM matrix approach used and encourage others to follow this example. Beyond that, for our purposes, we focused on what is considered convergent validity within the MTMM matrix approach (Raykov 2011) to help establish criterion validity. However, depending on the construct(s) under consideration in future research, it may be illustrative to consider issues of discriminant validity, which can also be effectively incorporated into the MTMM matrix approach. It is important to recognize that to be as defensible as possible scholars should consider providing validity evidence based on a series of related (and potentially unrelated) constructs to ensure an accurate understanding of the nomological network around a given single-item measure relative to a multi-item measure for the same construct.

An important caveat here is that, again, other validity data may be more suitable to present for a given research effort. For example, it may be informative to consider issues of predictive validity for a given construct if said construct is conceptually dynamic; demonstrating that a focal construct differentially predicts some relevant outcome based on different temporal lags may be informative. What is key here is that researchers proactively consider the types of validity evidence that presents a compelling case for a given construct and resulting measure. By extension then, we would discourage scholars from assuming some prescriptive sets of evidence are required when validating single-item measures or falling into the trap of thinking that the same evidence should be provided as compared to validating a multi-item measure.

Practical Implications

Various constraints (e.g., budget, time, sample access) and academic incentive structures (e.g., assessing focal

constructs with multi-item measures) drive many researchers to create surveys that are long and repetitive, despite the known negative impact of these survey features on representation and measurement (Čehovin et al. 2018). When responding to long, seemingly redundant surveys a substantial proportion of respondents are known to prematurely quit (creating non-response error) or invest little effort (creating psychometric detriments due to satisficing, careless responding, or insufficient effort responding; Callegaro et al. 2015; Gibson and Bowling 2020). Put simply, as participants become exhausted or cognitively drained, they are likely to begin responding differently and carelessly (e.g., Tourangeau 2018). Demonstrative of the incidence rate of careless responding, Bowling et al. (2021) estimate that for an online survey with 117 items, careless responding would occur 10% of the time, compared to just 1% of the time for an online survey with 33 items.

A well-accepted norm within organizational sciences is to deal with these threats after data collection, such as acknowledging survey representativeness as a limitation and statistically detecting or screening problematic responders (e.g., Meade and Craig 2012). Recommended interventions to ameliorate the negative effects of long surveys involve actions taken *during* survey administration such as warnings, interactive prompts, and in-person proctoring (Bowling et al. 2021). Echoing that of others, one overlooked straightforward solution to address issues of careless responding is to explore the use of abbreviated measures (Heggestad et al. 2019) or single-item indicators (Furnham 2008). Shifting from the dominant reactive paradigm (e.g., screening careless responders on long surveys) to a more proactive approach (e.g., shortening survey length through uptake and acceptance of single-item measures) is well aligned with increased efforts to define various best practices in research (e.g., Aguinis et al., 2021). This compendium serves as a critical resource to support that shift, making the reality of delivering short, user-friendly surveys possible while meeting the expectations in academic publishing to rely on previously validated measurement tools.

Beyond addressing issues like survey non-response, breakoff, and careless responding, leveraging single-item measures has the potential to help organizational scholars proactively address the ever-pervasive research-practice gap. That is, it is well recognized that for academics looking to collaborate with organizations, presenting a seemingly redundant survey (i.e., consisting of a series of multi-measures) to organizational stakeholders is likely to result in challenges and may even endanger the collaboration (Lapierre et al., 2018). Normalizing the use of single-item measures within scholarly programs of research may facilitate opportunities and collaborations with organizations reticent to administer long "ivory tower" surveys.

Limitations and Additional Directions for Future Research

As with any scholarly endeavor, our research must be evaluated relative to its limitations. First, while a systematic review of the literature was conducted, wherein we included a host of constructs, many more constructs were excluded based on a series of inclusion and exclusion criteria (e.g., a multi-item measure of the same construct should exist in the literature; an item should be able to be interpreted in different temporal windows). By design, then, we focused on constructs more likely to result in valid single-item measures. In doing so though, our results may overstate the application of single-item measures. We recommend researchers continue to develop new single-item measures of constructs not included here.

Again, we want to stress, our research should not be interpreted to suggest that single-item measures are a panacea, nor that all single-item measures are "valid." While this program of research provides validity evidence for the constructs under consideration, as the application of single-item continues, it will still be incumbent on researchers (a) to provide validity evidence for these and new single-item measures as continuous validation efforts (Flake 2021) as well as (b) to provide coherent justifications why using the single-item in the given study context is appropriate. It is common for researchers to use a psychological measure, whether single or multi-item, *without* accumulating and/or reporting strong validity evidence to support its interpretation and use in (new) contexts, populations, and/or applications (e.g., Flake 2021; Flake et al. 2017). By committing to robust initial and continuous validation efforts, single-item measure developers and users will play an important role in combatting this concerning measurement trend (Allen et al. 2022). Also, there may be a context where using a single-item measure is not appropriate. For example, using a single-item measure of *conscientiousness* (assuming it has been developed) in the selection context in order to select one of job applicants may not be appropriate. Providing a sound rationale for why the single-item measure in the given study context is appropriate will play a crucial role in facilitating scholars' acceptance of the use of single-item in the organizational psychology research.

In turn, similar to work examining factors that predict measure reliability (e.g., Tourangeau et al. 2020), theoretical rationale for why single-item measures for some constructs demonstrated higher construct validity compared to others needs to be developed. Our findings provide some insight into the reasons why, such that, for example, content validity of multi-item measures indirectly influences consistency-based reliability of single-item measures for the same construct. However, additional research is recommended to extend our understanding of which characteristics of the construct itself (e.g., job attitude vs job

characteristic) may influence the reliability and validity of single-item measures.

Conclusion

The knee-jerk reaction that *all* single-item measures imply a weak research design is counterproductive and serves to limit advancements in the organizational sciences. While there are constructs where single-item measures may not be appropriate, our research makes clear that it is possible to develop measures that accurately and reliably represent a given construct. In light of the practical advantages afforded by their use, we encourage researchers to proactively consider how leveraging single-item measures may help address existing and emerging conceptual, methodological, and empirical challenges within their given research domain.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10869-022-09813-3.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Aguinis, H., Hill, N. S., & Bailey, J. R. (2021). Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods, 24*, 678–693.

Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science. *European Journal of Psychological Assessment, 38*(1), 1–5.

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*(5), 732–740.

Ang, L., & Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research, 58*(2), 218–227.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Egeland, T. (2018). The failing measurement of attitudes: How semantic determinants of individual survey responses come to replace measures of attitude strength. *Behavior Research Methods, 50*(6), 2345–2365.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44*(2), 175–184.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin, 143*(3), 263–292.

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods, 24*, 718–738.

Boyd, B. K., Gove, S., & Hitt, M. A. (2005). Construct measurement in strategic management research: Illusion or reality? *Strategic Management Journal, 26*(3), 239–257.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 12*(3), 291–294.

Callegaro, M., Manfreda, K. L., & Vehovar, V. (2015). *Web survey methodology*. Sage.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Casper, W. J., Vaziri, H., Wayne, J. H., DeHauw, S., & Greenhaus, J. (2018). The jingle-jangle of work–nonwork balance: A comprehensive and meta-analytic review of its meaning and measurement. *Journal of Applied Psychology, 103*(2), 182–214.

Čehovin, G., Bosnjak, M., & Lozar Manfreda, K. (2018). Meta-analyses in survey methodology: A systematic review. *Public Opinion Quarterly, 82*(4), 641–660.

Cheah, J. H., Sarstedt, M., Ringle, C. M., Ramayah, T., & Ting, H. (2018). Convergent validity assessment of formatively measured constructs in PLS-SEM. *International Journal of Contemporary Hospitality Management, 30*(11), 3192–3210.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology, 104*(10), 1243–1265.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates, Inc.

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., & Banks, G. C. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology, 105*(12), 1351–1381.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *Teachers of English to Speakers of Other Languages Quarterly, 42*(3), 475–493.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.

De Vries, R. E., Realo, A., & Allik, J. (2016). Using personality item characteristics to predict single-item internal reliability, retest reliability, and self-other agreement. *European Journal of Personality, 30*(6), 618–636.

Dormann, C., & Van de Ven, B. (2014). Timing in methods for studying psychosocial factors at work. In M. Dollard, A. Shimazu, R. B. Nordin, P. Brough, & M. Tuckey (Eds.), *Psychosocial factors at work in the Asia Pacific* (pp. 89–116). Springer.

Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research, 3*(3), 196–204.

DuBay, W. H. (2004). *The principles of readability*. Impact Information, Costa Mesa, CA.

Edwards, J. R. (2003). Construct validation in organizational behavior research. In J. Greenberg (Ed.), *Organizational behavior: The state of the science* (pp. 327–371). Lawrence Erlbaum Associates.

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*(1), 3–23.

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist, 56*(2), 132–141. https://doi.org/10.1080/00461520.2021.1898962

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.

Fowler Jr., F. J. & Cosenza, C. (2009). Design and evaluation of survey questions. In *The SAGE handbook of applied social research methods* (pp. 375–412). SAGE Publications, Inc.

Fowler, F. J. (1995). *Improving survey questions*. Sage.

Freedle, R. (2003). Correcting the SAT's ethnic and social-class bias: A method for reestimating SAT scores. *Harvard Educational Review, 73*(1), 1–44.

Freedle, R., & Kostin, I. (1992). *The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: Main ideas, inferences and explicit statements (GRE Board Professional Report 87–10P; ETS RR-91-59)*. Educational Testing Service.

Freedle, R., & Kostin, I. (1997). Predicting Black and White differential item functioning in verbal analogy performance. *Intelligence, 24*(3), 417–444.

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft, 69*(2), 195–210.

Furnham, A. (2008). Relationship among four Big Five measures of different length. *Psychological Reports, 102*(1), 312–316.

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology, 15*(4), 380–387.

Gibson, A. M., & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment, 36*(2), 410–420.

Gilbert, S., & Kelloway, E. K. (2014). Using single items to measure job stressors. *International Journal of Workplace Health Management, 7*(3), 186–199.

Göritz, A. S (2014). Determinants of the starting rate and the completion rate in online panel studies. In M. Callegaro, R., Baker, J., Bethlehem, A. S., Göritz, J. A., Krosnick, & P. J. Lavrakas (Eds.) *Online panel research: A data quality perspective* (pp. 154–170). Wiley.

Guion, R. (1965). *Personnel testing*. McGraw-Hill.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items* (1st ed.). Routledge.

Hayes, A. F. (2018). Partial, conditional, and moderated moderated mediation: Quantification, inference, and interpretation. *Communication Monographs, 85*(1), 4–40.

Heggestad, E. D., Scheaf, D. J., Banks, G. C., Monroe Hausfeld, M., Tonidandel, S., & Williams, E. B. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management, 45*(6), 2596–2627.

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

Hinkin, T. R., & Tracey, J. B. (1999). An analysis of variance approach to content validation. *Organizational Research Methods, 2*(2), 175–186.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*(1), 99–114.

Hughes, D. J. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary approach to survey, scale and test development*. Chichester, UK: Wiley.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel*, Research Branch Report 8–75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163.

Lapierre, L. M., Matthews, R. A., Eby, L. T., Truxillo, D. M., Johnson, R. E., & Major, D. A. (2018; Focal Article). Recommended practices for initiating and managing research partnerships with organizations. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *11*(4), 543–581.

Lian, H., Ferris, D. L., & Brown, D. J. (2012). Does power distance exacerbate or mitigate the effects of abusive supervision? It depends on the outcome. *Journal of Applied Psychology, 97*(1), 107–123.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69.

McDonald, R. P. (1999). *Test theory: A unified approach*. Erlbaum.

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology, 65*(2), 287–321.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Muthén, L. K., & Muthén, B. O. (2018). *Mplus user's guide* (8th ed.).

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology, 75*(1), 77–86.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17*(4), 372–411.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw- Hill.

Oakland, T., & Lane, H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4*(3), 239–252.

Olson, K. (2008). Double-barreled question. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods*. Sage.

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17*(6), 609–626.

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research, 16*(1), 6–17.

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*(5), 879–903.

Raykov, T. (2011). Evaluation of convergent and discriminant validity with multitrait–multimethod correlations. *British Journal of Mathematical and Statistical Psychology, 64*(1), 38–52.

Robinson, S. B., & Leonard, K. F. (2018). *Designing quality survey questions*. Sage.
Rogelberg, S. G., & Stanton, J. M. (2007). Introduction: Understanding and dealing with organizational survey nonresponse. *Organizational Research Methods, 10*(2), 195–209.
Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing, 19*(4), 305–335.
Rossiter, J. R. (2008). Content validity of measures of abstract constructs in management and organizational research. *British Journal of Management, 19*, 380–388.
Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4*(1), 61–79.
Sarstedt, M., & Wilczynski, P. (2009). More for less? A comparison of single-item and multi-item measures. *Die Betriebswirtschaft, 69*(2), 211–227.
Scarpello, V., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology, 36*(3), 577–600.
Scherbaum, C. A., & Goldstein, H. W. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement, 68*(4), 537–553.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353.
Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management, 19*(2), 385–417.
Singh, J. (2003). A reviewer's gold. *Journal of the Academy of Marketing Science, 31*(3), 331–336.
Slaney, K. L., & Garcia, D. A. (2015). Constructing psychological objects: The rhetoric of constructs. *Journal of Theoretical and Philosophical Psychology, 35*(4), 244–259.
Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102–111.
Spector, P. E., Rosen, C. C., Richardson, H. A., Williams, L. J., & Johnson, R. E. (2017). A new perspective on method variance: A measure-centric approach. *Journal of Management, 45*(3), 855–880.
Spörrle, M., & Bekk, M. (2014). Meta-analytic guidelines for evaluating single-item reliabilities of personality instruments. *Assessment, 21*(3), 272–285.
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). Lawrence Erlbaum Associates.
Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.
Thurstone, L. L. (1947). *Multiple factor analysis*. University of Chicago Press.
Tourangeau, R. (2018). The survey response process from a cognitive viewpoint. *Quality Assurance in Education, 26*(2), 169–181.
Tourangeau, R., Yan, T., & Sun, H. (2020). Who can you count on? Understanding the determinants of reliability. *Journal of Survey Statistics and Methodology, 8*, 903–931.
Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods, 4*(4), 361–375.
Wanous, J. P., & Reichers, A. E. (1996). Estimating the reliability of a single-item measure. *Psychological Reports, 78*(2), 631–634.
Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82*(2), 247–252.
Wayne, J. H., Michel, J., & Matthews, R. A. (2022). Balancing work and family: A theoretical explanation and longitudinal examination of its relation to spillover and role functioning. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0001007
Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality, 44*(2), 180–198.
Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales-Five misunderstandings and ways to overcome them. *Journal of Individual Differences, 35*(4), 185–189.