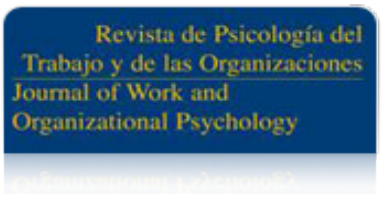

Revista de Psicología del Trabajo y de las Organizaciones
ISSN: 1576-5962
ISSN: 2174-0534
Colegio Oficial de la Psicología de Madrid

Golubovich, Juliya; Lake, Christopher J.; Anguiano-Carrasco, Cristina; Seybert, Jacob
Measuring Achievement Striving via a Situational Judgment Test: The Value of Additional Context
Revista de Psicología del Trabajo y de las Organizaciones, vol. 36, no. 2, 2020, pp. 157-168
Colegio Oficial de la Psicología de Madrid

DOI: https://doi.org/10.5093/jwop2020a15

Available in: http://www.redalyc.org/articulo.oa?id=231364008

Revista de Psicología del Trabajo y de las Organizaciones
ISSN: 1576-5962
ISSN: 2174-0534
Colegio Oficial de la Psicología de Madrid

Measuring Achievement Striving via a Situational Judgment Test: The Value of Additional Context

Golubovich, Juliya; Lake, Christopher J.; Anguiano-Carrasco, Cristina; Seybert, Jacob
Measuring Achievement Striving via a Situational Judgment Test: The Value of Additional Context
Revista de Psicología del Trabajo y de las Organizaciones, vol. 36, no. 2, 2020
Colegio Oficial de la Psicología de Madrid
Available in: http://www.redalyc.org/articulo.oa?id=231364008008
DOI: 10.5093/jwop2020a15

research-article

Measuring Achievement Striving via a Situational Judgment Test: The Value of Additional Context

Medición del esfuerzo hacia el logro mediante de un test de juicio situacional: el valor del contexto adicional

Juliya Golubovich [a*]
Indeed, Princeton Junction, New Jersey, USA
Christopher J. Lake [b]
Kansas State University, USA
Cristina Anguiano-Carrasco [c]
ACT, Lawrenceville, New Jersey, USA
Jacob Seybert [d]
Imbellus, USA

Revista de Psicología del Trabajo y de las Organizaciones, vol. 36, no. 2, 2020

Colegio Oficial de la Psicología de Madrid

Received: 28 September 2019
Accepted: 25 May 2020

DOI: 10.5093/jwop2020a15

ABSTRACT: The study extends personality and situational judgment test (SJT) research by using an SJT to measure achievement striving in a contextualized manner. Employed students responded to the achievement striving SJT, traditional personality scales, and workplace performance measures. The SJT was internally consistent, items loaded on a single factor, and scores converged with other measures of achievement striving. The SJT provided incremental criterion-related validity for the performance criteria beyond less-contextualized achievement striving measures. Findings suggest that achievement-related work scenarios may provide additional criterion-relevant information not captured by measures that are less contextualized.

Keywords: Situational judgment tests, Achievement striving, Personality traits, Interactionism, Contextualization, Bandwidth-fidelity.

RESUMEN: Este estudio extiende la investigación sobre los tests de juicio situacional (TJS) y la personalidad, usando un TJS para medir el esfuerzo hacia el logro de una manera contextualizada. Estudiantes con empleo respondieron el TJS de esfuerzo hacia el logro, escalas típicas de personalidad y medidas de desempeño en el trabajo. El TJS fue internamente consistente, los ítems cargaron en un único factor y las puntuaciones convergieron con otras medidas de esfuerzo hacia el logro. El TJS añadió validez relacionada con el criterio, para los criterios de desempeño, a la validez de las medidas menos contextualizadas de esfuerzo hacia el logro. Los hallazgos sugieren que escenarios de trabajo relacionados con el logro pueden proporcionar información adicional relevante para el criterio no capturada por medidas menos contextualizadas.

Palabras clave: Tests de Juicio Situacional, Esfuerzo hacia el logro, Rasgos de personalidad, Interaccionismo, Contextualización, Amplitud - Fidelidad.

Introduction

Personality has been a frequently studied subject in the area of employee selection and workplace behavior (see Sackett et al., 2017). Personality tests have been shown to predict important work-related criteria (e.g., training performance, job performance, wages; Almlund et al.,

2011; Barrick et al., 2001). Conscientiousness is consistently useful for predicting "typical" or "will-do" performance across jobs; the importance of the other Big Five factors is more variable across job types and criteria (Barrick et al., 2001; Judge et al., 2013; Sackett & Walmsley, 2014; Salgado & Táuriz, 2014; Wilmot & Ones, 2019).

Given frequently reported modest associations of self-reported personality measures with performance outcomes (e.g., typically maxing out in the .20s range for conscientiousness and lower for other Big Five traits; see Sackett & Walmsley, 2014 for an overview) and self-reported personality measures' susceptibility to faking, optimization of personality measures is a topic of ongoing interest (Kasten et al., 2018; Sackett et al., 2017). For example, researchers have examined the effects of different scoring approaches (e.g., single-statement vs forced-choice) on personality tests' validity (e.g., Fisher et al., 2019; Salgado & Táuriz, 2014) and susceptibility to faking (e.g., Cao & Drasgow, 2019).

A guiding principle for improving prediction accuracy is to attend closely to the nature of the performance criterion and to use this information to inform the choice or design of predictor measures (Hogan & Roberts, 1996). Given that individuals do not necessarily behave in the same ways across different life domains (e.g., at home, at work), it is appropriate to reference workplace situations in personality measures that are meant to predict behavior in the workplace. Indeed, personality measures that refer specifically to the workplace can yield greater validity than personality measures that ask respondents to indicate their general behavioral tendencies (Shaffer & Postlethwaite, 2012). Additionally, researchers have suggested that depending on the complexity versus specificity of the criterion, broader or narrower traits may prove more valid predictors. Prediction of relatively specific criteria may be enhanced by using facet-level, rather than broader traits (e.g., Griffin & Hesketh, 2004; Hough, 1992; Woo et al., 2015). With regard to the goal of reducing faking, applying situational judgment tests (SJTs) to the measurement of personality has shown some promise (Kasten et al., 2018).

In the current paper, we examine the validity of an SJT designed to assess the achievement striving facet of conscientiousness via workplace situations. Achievement striving (typically referred to as "achievement" in this paper for brevity) describes behaviors associated with working toward goals and other positive outcomes. Individuals who are high in this trait can be described as hard working, ambitious, confident, and resourceful (Drasgow et al., 2012). They may prioritize their work-related objectives over their personal lives (Neuman & Kickul, 1998). Achievement striving is one of the most criterion-valid facets of conscientiousness (Dudley et al., 2006). We examine our assessment's criterion-related validity for a set of relatively specific performance criteria that we expect to be influenced by achievement striving, and examine its incremental validity beyond less-contextualized measures of the same trait. This research contributes to the existing literature by (1) investigating the value of using contextualized personality facet measures

to predict work performance and (2) evaluating the amenability of the SJT format to the measurement of personality facets. To our knowledge, there are no published studies examining whether achievement striving can be reliably and validly measured with an SJT.

Next, we summarize existing research on the implications of personality measure contextualization and predictor-criterion matching for test validity. We also discuss the traditionally non-construct-focused approaches to SJT development and highlight the potential benefits of construct-focused strategies.

Contextualized Measurement of Personality

Human behavior can be viewed as a function of individuals' traits and psychologically active situational features (Funder, 2006; Tett & Burnett, 2003). Psychologically active features "trigger particular cognitive and affective processes that ultimately lead to predictable responses in feelings, thoughts, and actions" (Zayas et al., 2008, p. 378). According to the Cognitive Affective Processing System (CAPS) theory of personality, situations characterized by different psychologically active features can elicit different responses from the same individual (Mischel & Shoda, 1995). Relatedly, Trait Activation Theory (TAT; Tett & Burnett, 2003) suggests that situations differ in the extent to which they cue the behavioral expression of various traits. Overall, these theories suggest that understanding situations is integral to understanding individuals' behavior.

Applying these ideas to personality measurement, researchers have found that asking people to respond to items that reference a specific context can increase validity (see Shaffer & Postlethwaite, 2012). Lievens and Sackett (2017) define contextualization as "the extent to which stimuli are embedded in a detailed and realistic context" (p. 51). Using Lievens and Sackett's (2017) framework, personality scale items that ask individuals to describe themselves in general are considered decontextualized since they provide no contextual cues. Items that ask about behavior in certain types of situations, such as "at work," provide a low level of context. Situational judgment tests like the one developed here, have been described as providing a medium level of context because they ask individuals to predict their behavior given a context described in terms like "who," "when," "where," and "why" (Lievens & Sackett, 2017).

Applying low levels of contextualization to otherwise decontextualized broad and narrow personality measures has been shown to improve prediction of criteria (e.g., Lievens, De Corte, et al., 2008; Wang & Bowling, 2016; Woo et al., 2015). Lievens, De Corte, et al. (2008) and Woo et al. (2015) showed that contextualizing the achievement facet of conscientiousness this way can result in validity gains when predicting college students' academic performance.

Matching the Bandwidth of Predictors and Criteria

In addition to measuring personality in a more contextualized fashion, researchers have suggested that validity should be enhanced via appropriate matching of predictors to criteria on bandwidth. When criteria are relatively broad or multidimensional (e.g., overall job performance), predictors should be similarly broad to effectively map the criterion space, but fidelity (accuracy) of measurement is sacrificed (Hogan & Roberts, 1996). When a criterion is narrower (e.g., specific performance dimension), more homogenous predictors may provide better fidelity. Further, different facets of a trait may show differential relationships with a criterion, and, as a result, the overall trait score will relate more weakly to that criterion than do some of the facet scores (Hastings & O'Neill, 2009). Some meta-analytic findings support the strategy of matching predictors and criteria on bandwidth by showing that facet personality measures can enhance the prediction of narrow criteria (Dudley et al., 2006). However, this issue is still debated as more recent research suggests that facet-level measures may not provide incremental value when their unique variance is isolated from that of the Big Five traits (e.g., Salgado et al., 2015). Researchers have encouraged further investigation of the value of contextualized measures of personality and the conditions under which facet-level scales may outpredict factor-level scales (e.g., Ashton et al., 2014; Lievens, 2017; Shaffer & Postlethwaite, 2012).

Approaches to SJT Development

When the criteria of interest are relatively narrow, asking respondents to respond to specific work situations (medium level of context), as we do in the current study, may be more informative than having respondents consider how they behave in general (i.e., no context) or how they behave "at work" (i.e., low level of context). SJTs commonly require test takers to indicate the "best," or their preferred ("most likely") ways to respond to hypothetical work-related situations presented via detailed prompts. As such, SJTs are well-suited for assessing personality in a contextualized manner (Campion & Ployhart, 2013; Lievens, 2017).

SJTs have traditionally been developed by sampling situations ("critical incidents") from a target job. Job experts (e.g., employees, managers) help generate possible behavioral response options for a given scenario and test takers' responses to the options are scored relative to responses job experts believe to be most (and/or least) appropriate. This development approach contributes to SJTs' criterion-related validity, ability to provide applicants with a realistic job preview, and consequently, their popularity as selection tools (Christian et al., 2010; Lievens & Sackett, 2012). The expectation that SJTs are also less susceptible to faking than traditional personality measures also adds to their appeal (Kasten et al., 2018). However, the predominantly non-construct-centered approach to SJT development—as the SJT scenarios and behavioral response options

generated by job experts generally do not map to specific constructs—does not lend itself to strong construct-related validity (Christian et al., 2010) or adherence to standards for internal consistency reliability (e.g., Lievens, Peeters, et al., 2008). 1 Lack of construct clarity is a commonly noted limitation of the SJT methodology (e.g., Lievens & Motowidlo, 2016).

Among other benefits, construct-focused SJTs (relative to SJTs designed as samples of job performance) may have better internal consistency, clearer factor structures, and better generalizability across jobs. They may also prove more amenable to test improvement through construct refinement, and enable better construction of test batteries where each assessment provides incremental validity in predicting targeted criteria (by assessing additional relevant constructs; e.g., Beauregard, 2000; Gessner & Klimoski, 2006). Researchers have reported initial successes in using SJTs to measure personality traits (e.g., Arthur, 2017; Corstjens & Lievens, 2015; Kasten & Staufenbiel, 2015; Mussel et al., 2016) and other constructs (e.g., prosociality, emotion management; Motowidlo et al., 2016; Schlegel & Mortillaro, 2019). Arthur (2017) had test takers rate the effectiveness of SJT options that represented (dis)agreeable or (un)conscientious responses to the associated scenarios. He reported reliabilities of .75 (based on ratings of 15 response options) and .80 (based on ratings of 14 response options) for agreeableness and conscientiousness scores, respectively. The SJT scores correlated with their corresponding single-statement Likert-based measures (.31 for agreeableness, .28 for conscientiousness), and did not correlate with a measure of socially desirable responding. Kasten and Staufenbiel (2015) examined the susceptibility to faking of a 13-item conscientiousness SJT that required test takers to rate their level of agreement with response options that represented different levels of trait expression. The SJT was reliable (.76-.83 for the two studies) and although test takers were able to inflate their scores when asked to fake, the faking effect was significantly larger for a single-statement Likert-based measure of conscientiousness. Corstjens and Lievens (2015) used the AB5C model to create an SJT measure of agreeableness, conscientiousness, and extraversion, producing scores for positive and negative poles of these traits. Convergent correlations with Likert-based measures ranged from |.17| to |.36|. Conscientiousness scores explained 17% of the variance in students' peer-rated team assignment performance. Agreeableness and extraversion scores explained 1-2% of the variance in performance.

There is limited research into the feasibility of designing SJTs targeting personality facets. Mussel et al. (2016) designed SJT items for several personality facets, including self-consciousness, gregariousness, openness to ideas, compliance, and self-discipline. Test takers responded to 22 scenarios per trait, choosing from four response options (two representing a low level of the trait and two representing a high level). Reliabilities for self-consciousness, gregariousness, and openness to ideas were acceptable (.70-.75) but were lower for compliance and self-discipline (.55-.56). Convergent validities with corresponding single-

statement Likert-based measures ranged from .41 to .70, and average discriminant validity was -.01 (*SD* = 0.19). Yet, the SJT did not provide incremental validity for students' GPA over the decontextualized Likert-based facet measures. For a review of construct-driven SJTs more generally, please see Guenole et al. (2017).

Current Study

Given the limited research on SJT measures of personality facets, the ability of this methodology to assess facets in a reliable and valid manner needs further investigation. Thus, we aim to extend the recent research on personality SJTs to the achievement striving facet of conscientiousness. Given that the achievement SJT measure in the current study is new, we examine its construct validity and expect that:

*H*1: Achievement SJT scores will converge (i.e., show strong, positive correlations) with Likert-type measures of achievement and other facets of conscientiousness, and will be discriminant (i.e., show weaker correlations) from Likert-type measures of other personality constructs.

Relatedly, we examine the assessment's factor structure, and check whether a single-factor model will fit the data.

We chose to assess achievement striving because this facet of conscientiousness can be predictive of multiple dimensions of job performance. Job performance is generally considered to include both required and discretionary components. A common distinction made is that between task and citizenship performance (e.g., Rotundo & Sackett, 2002; Williams & Anderson, 1991). Task or "in-role" performance, involves contributing to the organization's production of goods or provision of services, and meeting the formal requirements of one's job (Rotundo & Sackett, 2002). Citizenship performance involves going above and beyond the minimum work requirements and is often further partitioned into behaviors directed toward the organization and other individuals in the workplace (e.g., Smith et al., 1983; Williams & Anderson, 1991).

Achievement striving has implications for both compulsory and discretionary work behaviors. Given their tendency to work hard and demonstrate their competence, individuals high in this trait are likely to want to quickly learn and master their primary work duties, resulting in relatively high levels of task performance. Because achievement striving also represents the desire to succeed, confidence, and resourcefulness, it should likewise affect individuals' tendency to exceed minimum work requirements and enact discretionary work behaviors. Consistent with earlier research (e.g., Chiaburu & Carpenter, 2013; Dudley et al., 2006; James, 1998), we expect that:

*H*2: Achievement SJT scores will be positively associated with task performance, citizenship toward the organization, and citizenship toward others.

A key goal of the current research is to evaluate the criterion-related validity benefits of additional contextualization of achievement

striving personality items. Based on research showing that even low contextualization of achievement items by adding the phrase "at school" to items provides incremental validity for performance in that same context over decontextualized measures of achievement (Lievens, De Corte et al., 2008; Woo et al., 2015) as well as meta-analytic findings on contextualization of personality items more broadly (Shaffer & Postlethwaite, 2012), we expect that:

*H*3: Achievement SJT scores will have incremental validity for predicting task performance, citizenship toward the organization, and citizenship toward others beyond both decontextualized and low-context Likert-type measures of achievement.

By leveraging the SJT method to create a contextualized measure of achievement striving, we extend research on the value of using contextualized personality facet measures to predict work performance and demonstrate the feasibility of developing construct-valid SJTs.

Method

Sample

Students. Employed U.S. college students participated in the study in exchange for partial course credit. The final sample size after cleaning, as described later, was 283. Participants were 67% female, 83% White, and 19.96 years old on average (*SD* = 3.13). Nearly everyone (98%) was employed at the time; the remainder had been employed within the past 18 months. For 82%, their job was/had been part-time (the remainder worked full-time—5%, seasonally—12%, or temporarily—1%). Jobs were most commonly in food preparation and serving (29%), sales (14%), office and administrative support (9%), and community and social services (9%). A few (18%) worked up to 10 hours per week in their jobs, 54% worked 11-20 hours, 16% worked 21-30 hours, 9% worked 31-40 hours, 3% worked 41-50 hours, and 0.4% worked 51-60 hours. Finally, 29% of participants consented to being rated by a supervisor; an additional 4% consented to being rated by a coworker. 2

Supervisors and coworkers. Of the final student sample, 26 were rated by a current/former supervisor and 6 were rated by a current/former coworker. The raters were mostly female (53%) and White (94%), and were 29.23 years old on average (*SD* = 9.89). Of the raters, 88% were working with the rated individual at the time of the survey; the remainder had worked with the ratee within the past year. The majority of the raters (72%) supervised or worked with the ratee between 3 and 18 months; for the remainder, the time span was longer (19%) or shorter (9%). Most had the opportunity to observe a ratee's performance several days a week (66%) or daily (16%). The vast majority were *certain* (28%) or *very certain* (63%) about the accuracy of their ratings.

Initial Development of the Achievement SJT

Following researchers' suggestion that personality SJTs could be developed by writing scenarios that elicit a particular trait and response options that span the trait continuum (e.g., Lievens &Motowidlo, 2016), we created a pool of 115 "items" (a scenario and its associated response options) to elicit achievement-related behaviors in the workplace. Researchers with training in psychology were provided with detailed instructions, including a definition of achievement and an existing pool of validated Likert achievement items that spanned the range of the construct. They wrote item stems expected to cue achievement striving (e.g., pursuing a promotion or some form of self-improvement, getting an assignment completed). For each stem they generated five behavioral response options that demonstrated varying levels of achievement striving. Each set of stems and response options was then peer reviewed by one or two researchers and, if needed, edited to enhance construct focus and differentiate response options.

To key the SJT items, we recruited raters on Mechanical Turk (MTurk) from among those over 18 who had at least 95% of their tasks accepted by other requesters. Though it is more common to use individuals with demonstrated expertise in an area for similar rating tasks, this sample was used due to sample size requirements associated with the large number of items that needed to be keyed. The resultant ratings were very similar to ratings obtained from a sample of people with training in I/O psychology. 3 Instructions oriented raters to the task by providing the same definition of achievement used by the item writers and illustrative behavioral examples associated with different achievement levels. They received a subset of SJT items from the pool of 110 and rated the response options on their construct level using a 7-point Likert scale ranging from 0 (*unrelated to achievement*) to 6 (*extremely high achievement*), with a 3 representing somewhat low achievement.

After data cleaning following recommended best practices (e.g., Meade & Craig, 2012) to ensure rating quality, each set of SJT response options associated with an item were rated on achievement by approximately 19 raters per item (*SD* = 2.33). Most of the raters (65%) had an Associate's degree or higher level of education and were employed outside of crowdsourcing sites at the time of the data collection (79%). The mean option ratings became the keyed point values. Response options with poor interrater agreement values (rwg < .50, using a uniform null distribution; LeBreton & Senter, 2008) were not keyed. (See Table 1 for a sample SJT item with point values). Five of the 115 items were dropped after this step because they had only two keyed options or the set of options did not appear to vary on achievement.

Next, we conducted an initial study of the SJT item pool with MTurk workers to examine individual items' correlations with measures of achievement (a traditional Likert-type measure), counterproductive work behavior, and organizational citizenship behavior. 4 Participants completed a random set of 15 SJT items along with the measures

described next. After data cleaning an average of 173 participants (*SD* = 7.04) completed each item in the 110-item pool. Based on these data, we selected 12 SJT items that showed the strongest correlations with the achievement and criterion measures, as well as the highest internal consistency when examining likelihood ratings of response options for a given SJT item.

Student Worker Measures

Achievement SJT. Participants completed the 12-item achievement SJT form (α = .84, calculated at the level of the 12 scored items), rating each response option using a six-point Likert scale ranging from 0 (*extremely unlikely* [to respond this way]) to 5 (*extremely likely* [to respond this way]). Likelihood judgments are considered more appropriate than effectiveness judgments when measuring personality and collecting responses to each option rather than asking for a single response per item helped maximize the amount of personality-relevant data gathered (Campion & Ployhart, 2013). These likelihood ratings were multiplied by the corresponding keyed achievement levels; likelihood ratings for low achievement response options (keyed at 3.0 or lower) were reverse-scored first so that higher values of the products would correspond to higher likelihood of high achievement-oriented behaviors. 5 Products were summed to calculate item scores for each participant. Item scores were averaged to calculate a total achievement SJT score.

Generic and workplace achievement striving. The generic (i.e., decontextualized) achievement striving scale consisted of 11 items (e.g., "Do more than what is expected of me"; α = .80) from Goldberg et al. (2006). 6 The workplace (low context) version of the achievement striving scale adapted the same 11 items to the workplace (e.g., "Do more than what is expected of me at my job"; α = .85). Participants responded to both measures using a five-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Item ratings were averaged to calculate scale scores for generic achievement striving and workplace achievement striving.

Personality. Big Five personality traits were measured using a 60-item scale from Soto and John (2017). Participants responded to the items using a five-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The scale yields scores for conscientiousness (e.g., "Is dependable, steady"; α = .87), extraversion (e.g., "Is full of energy"; α = .85), agreeableness (e.g., "Is helpful and unselfish with others"; α = .79), neuroticism (e.g., "Often feels sad"; α = .89), and openness (e.g., "Is complex, a deep thinker"; α = .84). The scale also yields scores for 15 personality facets, including sociability, assertiveness, energy level, compassion, respectfulness, trust, organization, productiveness, responsibility, anxiety, depression, emotional volatility, intellectual curiosity, aesthetic sensitivity, and creative imagination. Alpha values ranged from .52 to .83 with a median alpha of .73. Relatively low Cronbach's alpha coefficients, similar to those observed here, have

been noted by other researchers (Rammstedt et al., 2018; Soto & John, 2017); these coefficients may underestimate the measures' reliability (Rammstedt & Beierlein, 2014). Item ratings were averaged to calculate scale scores for the broad and narrow traits.

Table 1

Sample Achievement SJT Item

Rafi needs to finish a report for the company's senior leaders. However, he does not have the knowledge to write one of the sections. There are materials available that he could learn from but the report is due in two days. If you were Rafi, what would you do to finish the report? Please rate each option on how likely you would be to respond that way.	Keyed achievement level
a. Exclude the section you do not know how to write; work on it after you submit the report to have it ready if the senior leaders ask about it.	2.00
b. Ask senior leaders to extend your report deadline to have enough time to finish the difficult section.	2.73
c. Exclude the section you do not know how to write; chances are nobody will notice that it is missing.	1.00
d. Put aside some of your less important work, review your available sources, and write the difficult report section.	4.87
e. Cut back on your personal life and sleep over the next two days and do your best to include the difficult section in your report.	5.47

Note. Test takers respond using a Likert response scale ranging from extremely unlikely (0) to extremely likely (5).

Table 2

Sequential Data Cleaning Steps

Data exclusion criterion	Number of participants excluded	Rationale for exclusion
1. Answered more than one of the three attention check items incorrectly	49	Researchers (e.g., Meade & Craig, 2012) recommend using instructed response items to identify inattentive respondents. We allowed participants one incorrect response to account for potential miss-clicks. Others have reported proportions of inattentive respondents similar to ours (e.g., Ran et al., 2015).
2. Spent less than five minutes on the SJT measure	10	It is beneficial to pair special item-based screening (e.g., instructed response items) with post hoc data screening approaches like response time (e.g., Meade & Craig, 2012). Based on the number of words in the SJT and the average reading speed of 200 words/minute, the SJT items should have taken the average reader 10.6 minutes to read (not taking ratings into account). Five minutes was chosen as a conservative cut-off for identifying inattentive respondents.
3. Spent less than 15 seconds per item on more than one SJT item while also either failing the attention check embedded within the SJT or rating more than one SJT item with no variability across its set of response options	4	To identify respondents who perhaps began answering the SJT items attentively but subsequently had their attention decline, we implemented specific item time-based screening. Based on SJT item lengths and average reading speed, each item should have required the average reader 40.2-67.2 seconds to just read. Fifteen seconds per item was chosen as a conservative cut-off for identifying inattentive respondents for a given item. We allowed participants one short SJT item response time. In order to guard against screening out speedy readers who may have been responding attentively, we screened out individuals with short item response times only if they also failed the attention check within the item set or rated multiple SJT items with no variability.
4. Represented multivariate outliers based on a calculation of Mahalanobis distance performed on the item-level variables	3	Mahalanobis distance (Mahalanobis, 1936), which considers response patterns across a series of items and flags individuals furthest from the average response vector, is a recommended method for identifying inattentive respondents (Meade & Craig, 2012). A probability level of $p < .001$ was used to identify multivariate outliers.

Self-reported task performance. Task performance was measured using four items from Williams and Anderson (1991; α = .89). A sample item is "Adequately completed your assigned duties." Participants responded to the items using a five-point Likert scale ranging from 1 (*never*) to 5 (*very often*). Item ratings were averaged to calculate a self-reported task performance scale score.

Self-reported organizational citizenship. Organization-directed citizenship (OCBO) and individual-directed citizenship (OCBI) were each measured using 5 items from Podsakoff et al. (1990; α = .54 for OCBO; α = .79 for OCBI). Sample items were "Took extra breaks" (reversed; OCBO) and "Helped others who had been absent from work" (OCBI). Participants responded to the items using a five-point Likert scale ranging from 1 (*never*) to 5 (*very often*). After reversing the items as appropriate, item ratings were averaged to calculate OCBO and OCBI scale scores.

Background and employment information. Participants answered additional questions about their demographic background and employment.

Attention checks. Three selected response items were interspersed throughout the other measures to identify inattentive respondents. All

items instructed participants to choose a certain response (e.g., "For quality assurance purposes, please select *rarely* for this item").

Supervisor/Coworker Measures

Job performance. Supervisors and coworkers rated participants in this study along the same dimensions of job performance (using the same items but adjusting for other-ratings) as were used for participants' self-ratings: task performance (4 items; α = .91), OCBO (5 items; α = .71), and OCBI (5 items; α = .86). Item ratings were averaged to calculate scale scores.

In addition to responding to the same performance measures as the student participants, supervisors and coworkers completed two items rating participants' overall performance and promotability. Overall performance ("After considering everything you know about this employee, how would you rate his or her overall performance?") was rated using a five-point Likert scale ranging from 1 (*unsatisfactory*) to 5 (*outstanding*). Promotability ("This employee would be worthy of a promotion") was rated using a six-point Likert scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*). As these two items correlated at r = .80, we converted the ratings into z-scores and averaged them into a single score for overall performance and promotability (2 items; α = .87).

Background information. Supervisors and coworkers answered additional questions about their demographic background, nature and length of work relationship with ratee, and frequency of observation. Additionally, they rated their level of certainty regarding the ratings they provided on a 6-point Likert scale ranging from 1 (*very uncertain*) to 6 (*very certain*).

Procedure

Student workers completed the measures via an online survey. They were incentivized ($25 gift cards awarded by lottery) to volunteer contact information for a supervisor or coworker who had observed them at work. They previewed the questions others would answer to help them decide whether to volunteer someone's contact information. Supervisors/coworkers were invited to the study via email. They completed the online study measures after an initial screening to verify that they knew and had worked with the individuals they would rate.

Results

Data Cleaning

A total of 349 students participated in the study and completed the entire survey. Of these, 283 were retained for analyses. Table 2 summarizes the data cleaning steps to remove 66 apparently inattentive participants.

Relative to the included participants, the excluded participants evidenced lower: (a) time responding to the achievement SJT, $t(294.52) = -3.01$, $p < .01$; (b) achievement SJT scores, $t(347) = -13.97$, $p < .001$; (c) generic achievement striving scores, $t(81.86) = -6.74$, $p < .001$; (d) workplace achievement striving scores, $t(85.09) = -7.75$, $p < .001$; (e) self-rated task performance, $t(72.31) = -7.67$, $p < .001$; (f) self-rated OCBO, $t(75.71) = -7.93$, $p < .001$, and (g) self-rated OCBI, $t(347) = -3.50$, $p < .001$. Finally, a smaller percentage of excluded than included participants agreed to be rated, $\chi^2(1, N = 349) = 10.16$, $p < .001$. Overall, removing inattentive respondents somewhat restricted the ranges of key measures, which can attenuate effect sizes. However, including these respondents in our analyses did not appear to substantively change the study's findings.

Sampling Procedure Analyses

We compared the 95 participants who chose to provide contact information for a supervisor or coworker to the 188 who did not. Those who were willing to be rated self-reported higher levels of OCBO, $t(221.94) = 2.12$, $p < .05$. Scores on other measures did not differ significantly across the two subsamples. These results suggest that participants who perceived themselves as engaging in fewer citizenship behaviors toward the organization may have been reluctant to be rated and self-selected out from the second part of the study. Consequently, we may expect some restriction of range in the job performance ratings provided by supervisors and coworkers as compared to ranges that may have been observed if all study participants were rated.

Descriptive Statistics

Means, standard deviations, and correlations for the study variables are shown in Table 3. Lending convergent validity evidence to the achievement SJT, these scores were significantly correlated with generic achievement striving ($r = .37$, $p < .001$) and workplace achievement striving ($r = .43$, $p < .001$). Achievement SJT scores were also significantly correlated with dichotomized self-rated task performance ($r = .34$, $p < .001$), self-rated OCBO ($r = .43$, $p < .001$), and self-rated OCBI ($r = .33$, $p < .001$). A moderate correlation of achievement SJT scores with other-rated task performance ($r = .33$, $p = .06$) was also observed. Self-ratings of OCBO were significantly correlated with other-ratings of OCBO ($r = .46$, $p < .01$). Operational validities for the achievement SJT were estimated by correcting for criterion unreliability and indirect range restriction in achievement SJT scores. 7 The resulting validities were: .38 for dichotomized self-rated task performance, .61 for self-rated OCBO, .39 for self-rated OCBI, .37 for other-rated task performance, .21 for other-rated OCBO, and .33 for other-rated OCBI.

Table 3
Descriptives and Intercorrelations for Study Variables

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1. SJT	12.08	1.33	.84																																
2. GASS	4.19	0.47	.37*	.80																															
3. WASS	4.19	0.53	.43*	.72*	.85																														
4. Organization	3.96	0.85	.19*	.48*	.40*	.83																													
5. Productive	3.82	0.70	.36*	.60*	.52*	.63*	.71																												
6. Responsibility	3.79	0.68	.26*	.48*	.44*	.57*	.59*	.67																											
7. Sociability	3.30	0.94	.17*	.15*	.14*	.02	.16*	.04	.83																										
8. Assertiveness	3.43	0.88	.31*	.44*	.38*	.21*	.39*	.20*	.48*	.80																									
9. Energy level	3.85	0.66	.29*	.47*	.45*	.20*	.43*	.24*	.52*	.37*	.64																								
10. Compassion	4.00	0.65	.22*	.10	.16*	.14*	.12*	.25*	.09	-.06	.27*	.52																							
11. Respectful	4.19	0.60	.14*	.20*	.26*	.20*	.28*	.44*	.02	-.10	.21*	.49*	.66																						
12. Trust	3.55	0.70	.10	.10	.20*	.08	.16*	.22*	.05	-.04	.29*	.50*	.52*	.65																					
13. Anxiety	3.28	0.87	-.03	-.15*	-.20*	-.13*	-.27*	-.21*	-.21*	-.25*	-.25*	-.03	-.26*	-.28*	.75																				
14. Depression	2.42	0.89	-.20*	-.38*	-.34*	-.23*	-.41*	-.34*	-.36*	-.33*	-.52*	-.18*	-.31*	-.36*	.57*	.80																			
15. Volatility	2.49	0.90	-.13*	-.30*	-.30*	-.26*	-.39*	-.36*	-.12*	-.23*	-.24*	-.10	-.39*	-.32*	.63*	.59*	.81																		
16. Curiosity	3.98	0.63	.17*	.21*	.18*	-.04	.08	.10	-.03	.23*	.10	.14*	.08	.13*	-.04	-.08	-.12*	.59																	
17. Sensitivity	3.65	0.95	.02	-.01	-.02	-.11	-.16*	-.05	-.11	-.07	-.02	.13*	.12*	.19*	.01	.11	.04	.51*	.79																
18. Imagination	3.86	0.71	.18*	.26*	.21*	-.02	.11	.17*	.15*	.25*	.24*	.14*	.10	.12	-.09	-.15*	-.14*	.50*	.48*	.73															
19. C	3.86	0.63	.31*	.60*	.53*	.88*	.86*	.83*	.09	.31*	.33*	.20*	.35*	.17*	-.23*	-.37*	-.39*	.05	-.13*	.09	.87														
20. E	3.53	0.66	.31*	.42*	.38*	.17*	.39*	.19*	.86*	.79*	.74*	.10	.04	.10	-.29*	-.49*	-.24*	.12*	-.09	.26*	.29*	.85													
21. A	3.91	0.53	.18*	.16*	.25*	.17*	.22*	.36*	.07	-.08	.32*	.81*	.80*	.84*	-.23*	-.34*	-.33*	.14*	.18*	.14*	.28*	.10	.79												
22. N	2.73	0.76	-.14*	-.32*	-.33*	-.24*	-.42*	-.36*	-.27*	-.32*	-.39*	-.12*	-.37*	-.38*	.85*	.85*	.87*	-.09	.06	-.15*	-.39*	-.40*	-.35*	.89											
23. O	3.83	0.62	.13*	.16*	.13*	-.08	-.02	.07	-.01	.14*	.12	.16*	.13*	.18*	-.04	-.03	-.07	.79*	.86*	.79*	-.01	.10	.19*	-.06	.84										
24. SR task	0.71	0.45	.34*	.34*	.36*	.27*	.36*	.30*	.02	.12	.20*	.21*	.22*	.13*	-.05	-.19*	-.15	.06	.02	.04	.36*	.13*	.23*	-.15*	.04	.89									
25. SR OCBO	4.37	0.41	.43*	.43*	.49*	.30*	.40*	.45*	.07	.16*	.22*	.19*	.25*	.16*	-.06	-.20*	-.14*	.03	-.03	.12*	.44*	.18*	.24*	-.15*	.04	.41*	.54								
26. SR OCBI	4.16	0.65	.33*	.32*	.41*	.16*	.23*	.18*	.17*	.20*	.25*	.12*	.12	.21*	-.11	-.15*	-.12*	.10	.08	.20*	.22*	.25*	.18*	-.15*	.15*	.25*	.24*	.79							
27. OR task	4.70	0.46	.33	.36*	.36*	.49*	.33	.21	.13	.36*	.43*	-.05	.06	.16	-.10	-.30	-.12	.07	-.20	-.03	.40*	.39*	.08	-.20	-.09	.14	.47*	.02	.91						
28. OR OCBO	4.40	0.54	.16	.13	.18	.33	.14	.10	-.13	.00	.24	.10	.01	.14	.08	-.14	.07	.06	-.05	.00	.22	.04	.11	.00	.00	.28	.46*	-.11	.78*	.71				
29. OR OCBI	4.01	0.69	.29	.19	.39*	.36*	.36*	.10	.04	.20	.25	-.03	.07	.20	-.18	-.23	-.08	.33	.05	.26	.32	.22	.11	-.19	.24	-.07	.23	.15	.60*	.45*	.86			
30. OR overall	0.00	0.95	.26	.31	.42*	.57*	.40*	.26	.14	.38*	.31	.00	-.13	.01	-.19	-.23	-.15	.28	-.24	.18	.47*	.38*	-.04	-.22	.05	.02	.43*	.07	.76*	.60*	.81*	.87		
31. Job tenure	14.49	17.08	.06	-.06	.05	.00	.01	.02	-.01	.09	-.03	-.04	-.12	-.02	-.10	-.09	-.06	.07	.00	.05	.01	.02	-.07	-.10	.04	.01	-.05	.07	.07	-.01	.11	.09	-	
32. Gender	0.33	0.47	-.08	-.03	.01	-.18*	-.08	-.22*	.08	.16*	-.02	-.14*	-.18*	-.09	-.20*	-.07	-.13*	.19*	-.03	.14*	-.19*	.10	-.17*	-.16*	.10	-.16*	-.15*	.07	-.25	-.26	-.17	-.01	.13*	-
33. Age	19.96	3.13	-.01	.06	.09	.07	.08	.12*	-.10	.15*	-.08	.04	.00	-.02	-.10	-.03	-.09	.18*	.14*	.11	.10	-.01	.01	-.09	.18*	.06	.07	.07	.16	.16	.32	.32	.39*	.14*	-

Note. Ns = 31-33 for other-rated performance variables; Ns = 281-283 for remaining variables; GASS = generic achievement striving; WASS = workplace achievement striving; C = conscientiousness; E = extraversion; A = agreeableness; N = neuroticism; O = openness; SR = self-rated; OR = other-rated; OCBO = citizenship toward the organization; OCBI = citizenship toward other individuals; SR task performance was dichotomized; OR overall combines an overall performance and a promotability rating; Gender is coded 0 = female, 1 = male. Job tenure is in months; Diagonal values indicate scale reliabilities.
* p < .05 or better.

The convergence of these self- and other-ratings helps lend credibility to the self-reported performance data. Importantly, the small sample size for other-ratings ($N = 32$) meant low power for discerning statistically significant relations. Further, as we mentioned earlier, other-ratings may have some restriction of range. Therefore, the correlations involving other-reported performance variables generally may be under-estimated.

We should also note that relative to women, men rated their OCBO ($r = -.15$, $p < .05$) and task performance ($r = -.16$, $p < .05$) lower. Correlations of job tenure with the self- and other-ratings of performance were not statistically significant in this sample. Like earlier researchers (e.g., De Dreu & Nauta, 2009; Klotz et al., 2017), we included both gender and job tenure as control variables in our subsequent regression analyses.

Factor Analysis of SJT Items

Confirmatory factor analysis was performed on the achievement SJT items using Amos 25. We specified a model with 12 scored SJT items as observed variables and achievement as the latent variable. The model

appeared to have acceptable fit to the data, $\chi^2(54) = 120.59$, $p < .001$, RMSEA = .07, 90% CI [.05, .08], CFI = .92, TLI = .91.

Hypothesis Tests

Hypothesis 1. Hypothesis 1 predicted that achievement SJT scores will show strong, positive correlations with Likert-type measures of achievement striving and other facets of conscientiousness, and weaker correlations with Likert-type measures of other personality constructs. As mentioned previously and shown in Table 3, achievement SJT scores were moderately correlated with generic achievement striving ($r = .37$, $p < .001$) and workplace achievement striving ($r = .43$, $p < .001$). Achievement SJT scores were also moderately correlated with scores on conscientiousness's facets of productiveness ($r = .36$, $p < .001$), responsibility ($r = .26$, $p < .001$), and organization ($r = .19$, $p < .01$). This set of correlations lends convergent validity evidence to the achievement SJT.

With regard to discriminant validity, achievement SJT scores generally showed weaker correlations with extraversion's facet of sociability, and the facets of agreeableness, neuroticism, and openness (absolute *r* values ranging from .02 to .18). However, achievement SJT scores correlated relatively highly with extraversion's facets of assertiveness ($r = .31$, $p < 001$) and energy level ($r = .29$, $p < .001$), the compassion facet of agreeableness ($r = .22$, $p < .001$), and the depression facet of neuroticism ($r = -.20$, $p < .001$).

To formally test Hypothesis 1, we conducted pairwise comparisons of the correlations intended to show achievement SJT scores' convergent validity with those intended to show their discriminant validity. In 67% of the comparisons, the convergent validity correlations significantly (one-tailed according to the hypothesis) exceeded the discriminant validity correlations (z values ranging from 1.9, $p < .05$, to 5.83, $p < .001$; Lee & Preacher, 2013). In another 32% of the comparisons, the magnitudes of the convergent validity correlations did not differ significantly from the discriminant validity correlations. Finally, there was one case of a statistically significant difference counter to the hypothesis: achievement SJT scores correlated more highly with assertiveness ($r = .31$, $p < .001$) than with organization scores ($r = .19$, $p < .01$; $z = -1.65$, $p < .05$). Overall, Hypothesis 1 was partially supported. The relatively low correlation between achievement SJT scores and the organization facet of conscientiousness appeared to be the main reason that Hypothesis 1 did not receive stronger support.

Hypothesis 2. Hypothesis 2 predicted that achievement SJT scores would be positively associated with task performance, OCBO, and OCBI. SJT scores were directionally related to supervisors'/coworkers' task performance ($r = .33$), OCBO ($r = .16$), and OCBI ($r = .29$), but the $N = 32$ sample size resulted in little statistical power for hypothesis testing. To further examine these relationships, hierarchical regression analyses were conducted using the larger sample of self-reported versions

of these variables. Two ordinary least squares (OLS) regressions were used to examine the effects of achievement SJT scores on OCBO and OCBI after accounting for the effects of the control variables. As the self-reported task performance variable was highly negatively skewed (71% of participants rated themselves a "5" overall), we dichotomized these scores for a logistic regression. [8] Scores of 5 were treated as high ("1") and scores below 5 were treated as low ("0"). For both OLS and logistic regressions, gender and job tenure were entered in step 1, followed by the achievement SJT in step 2.

Regression results are shown in Tables 4 and 5. Beyond the control variables, achievement SJT scores explained significant amounts of additional variance in: task performance, $\chi^2(1, N = 281) = 31.36$, $p < .001$, Nagelkerke R^2 change = 14.7%; OCBO, $\Delta R^2 = 17.7\%$, $F(1, 277) = 61.34$, $p < .001$; and OCBI, $\Delta R^2 = 11.5\%$, $F(1, 277) = 36.50$, $p < .001$. Thus, bivariate correlations and regression analyses supported Hypothesis 2.

Table 4
Hierarchical Logistic Regressions for Self-Rated Task Performance

	SR task performance	
	B	Odds ratio
Hypotheses 2-3: step 1		
Gender	0.74	2.10
Job tenure	0.00	1.00
$\chi^2(2)$		7.19*
Nagelkerke R^2		0.04
Hypothesis 2: step 2		
Achievement SJT	0.61*	1.84
$\chi^2(1)$		31.36*
Nagelkerke R^2		0.18
Hypothesis 3: step 2		
Generic achievement striving	1.69*	5.44
$\chi^2(1)$		32.73*
Nagelkerke R^2		0.19
Hypothesis 3: step 3		
Workplace achievement striving	1.13*	3.09
$\chi^2(1)$		9.04*
Nagelkerke R^2		0.23
Hypothesis 3: step 4		
Achievement SJT	0.40*	1.49
$\chi^2(1)$		9.94*
Nagelkerke R^2		0.27

Note. SR = self-rated; for gender, 0 = male, 1 = female; Bs are unstandardized regression weights.
* p ≤ .05 or better.

Table 5
Hierarchical OLS Regressions for Self-Rated Citizenship toward the Organization, and Citizenship toward Other Individuals

Variables	SR OCBO		SR OCBI	
	r	β	*r*	β
Hypotheses 2-3: step 1				
Gender	-.15*	-.14*	.07	.06
Job tenure	-.05	-.03	.07	.06
ΔR^2		.02*		.01
Hypothesis 2: step 2				
Achievement SJT	.43*	.42*	.33*	.34*
ΔR^2		.18*		.12*
Hypothesis 3: step 2				
Generic achievement striving	.43*	.43*	.32*	.32*
ΔR^2		.18*		.10*
Hypothesis 3: step 3				
Workplace achievement striving	.49*	.38*	.41*	.37*
ΔR^2		.07*		.07*
Hypothesis 3: step 4				
Achievement SJT	.43*	.25*	.33*	.20*
ΔR^2		.05*		.03*

Note. β = standardized regression weight; OCBO = citizenship toward the organization; OCBI = citizenship toward other individuals; SR = self-rated.
* $p < .05$ or better.

Hypothesis 3. Hypothesis 3 predicted that achievement SJT scores would have incremental validity for predicting task performance, OCBO, and OCBI beyond both decontextualized and low context Likert measures of achievement. This hypothesis was also evaluated using hierarchical regressions. Gender and job tenure were entered in step 1, generic achievement striving in step 2, workplace achievement striving in step 3, and the achievement SJT in step 4.

Regression results are shown in Tables 4 and 5. Beyond the control variables and less-contextualized achievement measures, achievement SJT scores explained significant amounts of additional variance in: task performance, $\chi^2(1, N = 281) = 9.94$, $p < .01$, Nagelkerke R^2 change = 4.1%; OCBO, $\Delta R^2 = 4.9\%$, $F(1, 275) = 19.78$, $p < .001$; and OCBI, $\Delta R^2 = 3.1\%$, $F(1, 275) = 10.69$, $p < .01$. Thus, Hypothesis 3 was supported.

Discussion

There is ongoing interest in improving personality measurement for employee selection; contextualizing personality measures and assessing narrower traits when predicting relatively narrow criteria have been two effective approaches (Sackett et al., 2017). Our study contributes to research in this area by combining these approaches and taking contextualization a step further by assessing the achievement facet of

conscientiousness via an SJT. Unlike typical SJTs derived from critical incidents, the current construct-focused SJT showed good internal consistency reliability (on par with traditional Likert-type personality measures), items loaded reasonably well on a single achievement factor, and the SJT converged with other measures of achievement. Findings indicated practical value in applying the SJT method to measure achievement, as the achievement SJT explained incremental variance in task performance, OCBO, and OCBI beyond less-contextualized measures of achievement.

Theoretical Implications

Our study answers researchers' calls for continued investigation into contextualized measures of personality. This study also contributes to the ongoing debate about the value of contextualizing measures beyond just adding context tags (e.g., "at work") and untapped opportunities for cross-fertilization between the personality and industrial-organizational psychology fields (e.g., Lievens, 2017). Findings suggest that when measuring achievement, detailed work scenarios may provide additional criterion-relevant information not captured by measures that are lightly contextualized with "at work" tags. Christiansen and Speer (2017) suggested that lower bandwidth personality assessments may show reduced validity because they will capture responses to a narrow set of situational demands. As we sought alignment of our relatively narrow bandwidth achievement SJT to relatively narrow measures of job performance (i.e., performance dimensions as opposed to an overall performance composite), additional contextualization generally did not lead to the achievement SJT showing reduced validity. Ultimately, the criteria we examined are still fairly broad (Hogan & Roberts, 1996), so there would be value in investigating the validity of the achievement SJT relative to criteria that are even better matched to its bandwidth.

In addition to contributing to personality research by investigating and highlighting the value SJTs may bring to personality assessment, our findings contribute to SJT-related research by speaking to the feasibility of construct-focused SJT design. This work helps to fill a research gap, as studies examining SJT measures of personality facets have been scarce and no published studies have used an SJT to evaluate achievement striving. As discussed earlier, Mussel et al. (2016) designed SJT items for several personality facets, but their measure was intended for an academic, rather than a work, context. We extend their findings to an additional construct and context, similarly showing that a personality facet may be assessed reliably ($\alpha = .84$ in the current study) via an SJT and that the measure can demonstrate validity relative to criteria from the targeted context (in this case work).

Importantly, our findings indicate that achievement SJT scores, while showing promising convergent validity with other measures of achievement, may also reflect other traits that contribute to achievement striving behavior (e.g., assertiveness, energy level, compassion [in the

context of interpersonal situations]). In fact, these traits were likewise correlated with the traditional Likert-type achievement measures in our study. As has been pointed out by other researchers, SJTs may be inherently multidimensional due to their behavioral response options mirroring the complex nature of behavior in real situations, and this often leads to low internal consistency reliability and an uninterpretable factor structure (Catano et al., 2012; Lievens, Peeters, et al., 2008). In the current study, we were able to achieve favorable levels of reliability and unidimensionality (based on a confirmatory factor analysis) by developing scenarios expected to elicit achievement and creating response options that represented different levels of achievement. We concur with other researchers (e.g., Lievens &Motowidlo, 2016) that this type of test development strategy should be a viable approach for future construct-focused SJT development.

Practical Implications

By helping to address SJTs' typical lack of construct clarity, the current study can further promote the appeal of this methodology among both researchers and practitioners. For example, constructing test batteries with construct-focused SJTs should be easier than with SJTs that are based on critical incidents and thus are more heterogeneous in terms of the individual differences they capture (e.g., Gessner & Klimoski, 2006). Construct-based SJTs like this one should be applicable to a broad range of jobs, which is in contrast to typical SJTs that would be applicable only to jobs closely aligned to the role for which the SJT was developed (as the key represents judgments of experts about appropriate behavior in that role). The validity values in this study, based on participants who held a variety of jobs, support the broad generalizability of construct-based SJTs.

Importantly, an achievement SJT should not necessarily replace traditional personality inventories in a selection process. Measures of conscientiousness, in particular, are very useful for selecting individuals (e.g., a traditional, single-statement Likert-based measure of conscientiousness in this study correlated at .22-.44 with the self-reported performance criteria). Instead, an achievement SJT might be added to a selection battery to supplement existing measures. In order to be included in test batteries, an achievement SJT should explain incremental variance in job performance dimensions that companies care about. The decision to include a measure like this would need to factor in the extra cost associated with SJT development and the extra testing time associated with an additional measure. One of the strengths of the current SJT is that it consists of only 12 scenarios, but allows gathering 60 data points related to an individual's level of achievement by requiring ratings of all five response options per item. As we initially started with a larger pool of 110 items before winnowing the set down to a 12-item fixed form based on initial piloting work, we recommend that test developers plan to write more SJT items than they plan to use and subsequently select ones most likely to yield a construct-valid and reliable fixed form.

Future Research Directions

A limitation of the current study was the inability to use other-reported performance for hypothesis testing due to the small sample size for these ratings. We still briefly summarized our attempt to collect these ratings and the associated correlational results (correlations between SJT scores and other-rated performance were in the predicted direction), as these data seem instructive for better understanding the self-reported performance results and for highlighting which student workers tended to feel more comfortable getting evaluated. Our reliance on self-reported predictor and criterion variables may have introduced the potential for method bias (Conway & Lance, 2010). However, the observed convergence of OCBO self- and other-ratings was reassuring. We expect that additional correlations between self- and other-ratings may have been significant given a larger sample size of other-reports and if not for some likely restriction of range in other-ratings; student workers who perceived themselves as lower in OCBO were less likely to share contact information for a supervisor or coworker and raters less confident in their perceptions of the student workers may have self-selected out of the study. Further, research suggests that self-ratings of job performance can sometimes be as valid as other-ratings (Atwater et al., 1998). Regardless, it would be informative for generalizability purposes to examine the validity of the current SJT relative to either other-ratings of performance or more objective performance metrics (e.g., attendance, awards, promotions).

A second, related limitation had to do with some evidence of range restriction in achievement scores and the self-rated performance outcomes that we relied on for hypothesis testing. The student workers removed from analyses due to inattentive responding also appeared to be lower on the achievement trait. Further, participants tended to rate themselves fairly highly on task performance, requiring dichotomization of these ratings for analyses. We believe it was appropriate to remove inattentive respondents to enhance the internal validity of our study (Maniaci & Rogge, 2014), but this could have impacted the representativeness of our sample (Bowling et al., 2016). Reduced variability in key measures in this study may have actually somewhat attenuated true associations between these measures. The fact that our hypotheses were still supported may suggest the robustness of the achievement SJT validity-related findings. Nonetheless, further research in the context of a predictive validation study in an organization, where employees at various performance levels are included in the sample, would be a valuable way to examine the generalizability of current findings.

Some may view the use of university student participants as limiting the generalizability of this study's findings. But we would point to the diverse array of jobs, employment sectors, and employment relationships (e.g., full-time, part-time) that the participants represented. Further, we agree with those who suggest focusing more on the generalizability of the effect than the representativeness of the sample (Highhouse & Gillespie, 2009). It would be important to examine the generalizability of the effects

in this study to a higher-stakes selection context and to individuals who mostly work full-time.

Although we believe that the additional context provided by the SJT methodology contributed to the achievement SJT's incremental validity beyond decontextualized and low context measures of achievement, we are not able to rule out other explanatory mechanisms for these effects. For example, in contrast to the incremental validity provided by our achievement SJT, Mussel et al.'s (2016) SJT measures of personality facets generally did not provide incremental validity beyond corresponding traditional measures of the same traits for explaining variability in college GPA. Our items apparently included more context than theirs and correspondingly, were also longer. Researchers have suggested that SJT scores may reflect cognitive ability and that this may help to explain their incremental validity (e.g., Judge et al., 2017). Further, there is a current debate among researchers about the value of even including "the situation" in SJT items—implying that context may not matter (e.g., Krumm et al., 2015; Rockstuhl et al., 2015). In light of all these considerations, it is important for future research to clarify when and why construct-focused SJTs may improve personality measurement and prediction of criteria, as researchers' promotion of SJTs for personality assessment (e.g., Lievens, 2017) is at least partially predicated on SJTs' ability to do just that. For example, it would be valuable to investigate whether the current SJT has incremental validity beyond a measure of cognitive ability.

Finally, we echo Mussel et al.'s (2016) suggestion that it will be important for additional research to shed light on whether construct-focused SJTs still retain the desirable properties attributed to more typical, critical incidents-based SJTs, including their ability to provide applicants with a realistic job preview and lower susceptibility to faking relative to self-report measures of personality. A promising recent study by Kasten et al. (2018), discussed earlier, investigated the susceptibility to faking of an SJT designed to measure several Big Five traits, which, like our assessment, used response options that differed in their level of trait expression. They found the measure more resistant to faking than a corresponding traditional measure of personality. Similar research with achievement SJTs would be of value, especially if conducted in a high-stakes setting where individuals are motivated to "fake good."

Conclusion

We believe that construct-focused SJTs assessing personality facets hold great promise. In this study, we have shown that these SJTs can have many desirable properties, including strong internal consistency reliability, a more interpretable factor structure, strong validity, and broad generalizability to a variety of jobs. They can also offer additional predictive validity above and beyond other types of measures. Further research is recommended to better establish the extent to which such

SJTs' incremental validity beyond more traditional personality measures is due to extra contextual information.

References

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality psychology and economics* (No. w16822). National Bureau of Economic Research.

Arthur Jr., W. (April, 2017). Developing a construct-laden SJT of FFM traits: Ingenuity or folly? In J. Golubovich & C. Anguiano-Carrasco's (Co-chairs), *Development and scoring of construct-focused situational judgment tests*. Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology. Orlando, FL.

Ashton, M. C., Paunonen, S. V., & Lee, K. (2014). On the validity of narrow and broad personality traits: A response to Salgado, Moscoso, and Berges (2013). *Personality and Individual Differences, 56*, 24-28. https://doi.org/10.1016/j.paid.2013.08.019

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology, 51*, 577-598. https://doi.org/10.1111/j.1744-6570.1998.tb00252.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next?. *International Journal of Selection and Assessment, 9*(1-2), 9-30. https://doi.org/10.1111/1468-2389.00160

Beauregard, R. S. (2000). *Construct explication of a situational judgment test: Addressing multidimensionality through item development, content analysis, and scoring procedures* (Unpublished doctoral dissertation). Wright State University.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*, 218-229. https://doi.org/10.1037/pspp0000085

Campion, M. C., & Ployhart, R. E. (2013). Assessing personality with situational judgment measures: Interactionist psychology operationalized. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work* (pp. 439-456). Routledge.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*, 1347-1368. https://doi.org/10.1037/apl0000414

Catano, V. M., Brochu, A., & Lamerson, C. D. (2012). Assessing the reliability of situational judgment tests used in high-stakes situations. *International Journal of Selection and Assessment, 20*, 333-346. https://doi.org/10.1111/j.1468-2389.2012.00604.x

Chiaburu, D. S., & Carpenter, N. C. (2013). Employees' motivation for personal initiative: The joint influence of status and communion striving. *Journal of Personnel Psychology, 12*, 97-103. https://doi.org/10.1027/1866-5888/a000089

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related

validities. *Personnel Psychology, 63*, 83-117. https://doi.org/10.1111/j.1744-6570.2009.01163.x

Christiansen, N. D., & Speer, A. B. (2017). Putting situations into personality assessments: Problems and potential. *European Journal of Personality, 31*, 443-445.

Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business and Psychology, 25*, 325-334. https://doi.org/10.1007/s10869-010-9181-6

Corstjens, J., & Lievens, F. (April, 2015). Personality score variability across situations on SJTs: Inconsistency or flexibility? In M. Reeder and J. Golubovich's (Co-chairs), *Situational judgment test design and measurement informed by psychological theory*. Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology. Philadelphia, PA.

De Dreu, C. K. W., & Nauta, A. (2009). Self-interest and other-orientation in organizational behavior: Implications for job performance, prosocial behavior, and personal initiative. *Journal of Applied Psychology, 94*, 913-926. https://doi.org/10.1037/a0014494

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army selection and classification decisions* (United States Army Research Institute for the Behavioral and Social Sciences Technical Report No. 1311). U.S. Army Research Institute for the Behavioral and Social Sciences.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40-57. https://doi.org/10.1037/0021-9010.91.1.40

Fisher, P. A., Robie, C., Christiansen, N. D., Speer, A. B., & Schneider, L. (2019). Criterion-related validity of forced-choice personality measures: A cautionary note regarding Thurstonian IRT versus Classical Test Theory scoring. *Personnel Assessment and Decisions, 1*, 49-61. https://doi.org/10.25035/pad.2019.01.003

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality, 40*, 21-34. https://doi.org/10.1016/j.jrp.2005.08.003

Gessner, T. L., & Klimoski, R. J. (2006). Making sense of situations. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational Judgment Tests: Theory, measurement, and application* (pp. 13-38). Lawrence Erlbaum Associates, Inc.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96. https://doi.org/10.1016/j.jrp.2005.08.007

Griffin, B., & Hesketh, B. (2004). Why openness to experience is not a good predictor of job performance. *International Journal*

of Selection and Assessment, 12, 243-251. https://doi.org/10.1111/j.0965-075X.2004.278_1.x

Guenole, N., Chernyshenko, O. S., & Weekley, J. (2017). On designing construct driven situational judgment tests: Some preliminary recommendations. *International Journal of Testing, 17*, 234-252. https://doi.org/10.1080/15305058.2017.1297817

Hastings, S. E., & O'Neill, T. A. (2009). Predicting workplace deviance using broad versus narrow personality variables. *Personality and Individual Differences, 47*, 289-293. https://doi.org/10.1016/j.paid.2009.03.015

Highhouse, S., & Gillespie, J. Z. (2009). Do samples really matter that much? In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences* (pp. 247-266). Routledge.

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior, 17*, 627-637. https://doi.org/10.1002/(SICI)1099-1379(199611)17:6<627::AID-JOB2828>3.0.CO;2-F

Hough, L. M. (1992). The Big Five personality variables—Construct confusion: Description versus prediction. *Human Performance, 5*, 139-155. https://doi.org/10.1080/08959285.1992.9667929

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869-879. https://doi.org/10.1037/0021-9010.85.6.869

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods, 1*, 131-163. https://doi.org/10.1177/109442819812001

Judge, T. A., Hofmans, J., & Wille, B. (2017). Situational judgment tests and personality measurement: Some answers and more questions. *European Journal of Personality, 31*, 463-464. https://doi.org/10.1080/00062278.1970.10596758

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the Five-Factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*, 875-925. https://doi.org/10.1037/a0033901

Kasten, N., Freund, P. A., & Staufenbiel, T. (2018). "Sweet little lies": An in-depth analysis of faking behavior on situational judgment tests compared to personality questionnaires. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000479

Kasten, N., & Staufenbiel, T. (2015, May). A construct-oriented development approach to situational judgment tests. *Paper presented at the meeting of the European Association of Work and Organizational Psychology*. Oslo, Norway.

Klotz, A. C., Bolino, M. C., Song, H., & Stornelli, J. (2017). Examining the nature, causes, and consequences of profiles of organizational citizenship behavior. *Journal of Organizational Behavior, 39*, 629-647. https://doi.org/10.1002/job.2259

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests?

Journal of Applied Psychology, 100, 399-416. https://doi.org/10.1037/a0037674

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

Lee, I. A., & Preacher, K. J. (2013). *Calculation for the test of the difference between two dependent correlations with one variable in common* [Computer software]. http://quantpsy.org/corrtest/corrtest2.htm

Lievens, F. (2017). Integrating situational judgment tests and assessment centre exercises into personality research: Challenges and further opportunities. *European Journal of Personality, 31*, 441-502. https://doi.org/10.1002/per.2119

Lievens, F., De Corte, W., & Schollaert, E. (2008). *A closer look at the frame-of-reference effect in personality scale scores and validity*. *Journal of Applied Psychology, 93*, 268-279. https://doi.org/10.1037/0021-9010.93.2.268

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology, 9*, 3-22. https://doi.org/10.1017/iop.2015.71

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*, 426-441. https://doi.org/10.1108/00483480810877598

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

Lievens , F., & Sackett, P. R. (2017). The effects of predictor method factors on selection outcomes: A modular approach to personnel selection procedures. *Journal of Applied Psychology, 102*, 43-66. https://doi.org/10.1037/apl0000160

Mahalanobis, P. C. (1936). *On the generalized distance in statistics*. Proceedings of the National Institute of Sciences in India, 2, 49-55.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects of research. *Journal of Research in Personality, 48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

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

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246-268. https://doi.org/10.1037/0033-295X.102.2.246

Motowidlo, S. J., & Beier, M. E. (2010). Differentiating specific job knowledge from implicit trait policies in procedural knowledge measures by a situational judgment test. *Journal of Applied Psychology, 95*, 321-333.

Motowidlo, S. J., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance, 29*, 331-346. https://doi.org/10.1080/08959285.2016.1165227

Mussel, P., Gatzka, T., & Hewig, J. (2016, August 3). Situational judgment tests as an alternative measure for personality assessment. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000346

Neuman, G. A., & Kickul, J. R. (1998). Organizational citizenship behaviors: Achievement orientation and personality. *Journal of Business and Psychology, 13*, 263-279. https://doi.org/10.1023/A:1022963108025

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

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

Rammstedt, B., & Beierlein, C. (2014). Can't we make it any shorter? The limits of personality assessment and ways to overcome them. *Journal of Individual Differences, 35*, 212-220. https://doi.org/10.1027/1614-0001/a000141

Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2018, August). Validation of the short and extra-short forms of the Big Five Inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment*. Advance online publication. https://doi.org/10.1027/1015-5759/a000481

Ran, S., Liu, M., Marchiondo, L. A., & Huang, J. L. (2015). Difference in response effort across sample types: Perception or reality? *Industrial and Organizational Psychology, 8*, 202-208. https://doi.org/10.1017/iop.2015.26

Rockstuhl, T., Ang, S., Ng, K. Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology, 100*, 464-480. https://doi.org/10.1037/a0038098

Rotundo, M., & Sackett, P. R. (2002). The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach. *Journal of Applied Psychology, 87*(1), 66-80. https://doi.org/10.1037/0021-9010.87.1.66

Sackett, P. R., Lievens, F., Van Iddekinge, C. H., & Kuncel, N. R. (2017). Individual differences and their measurement: A review of 100 years of research. *Journal of Applied Psychology, 102*, 254-273. https://doi.org/10.1037/apl0000151

Sackett, P. R., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace? *Perspectives on Psychological Science, 9*, 538-551. https://doi.org/10.1177/1745691614543972

Sackett, P. P., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology, 85*(1), 112-118. https://doi.org/10.1037/0021-9010.85.1.112

Salgado, J. F., Moscoso, S., Sanchez, J. I., Alonso, P., Chorogwicka, B., & Berges, A. (2015). Validity of the five-factor model and their facets: The impact of performance measure and facet residualization on the bandwidth-fidelity dilemma. *European Journal of Work and Organizational Psychology, 24*, 325-349. https://doi.org/10.1080/1359432X.2014.903241

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories, and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*, 3-30. https://doi.org/10.1080/1359432X.2012.716198

Schlegel, K., & Mortillaro, M. (2019). The Geneva Emotional Competence Test (GECo): An ability measure of workplace emotional intelligence. *Journal of Applied Psychology, 104*, 559-580. https://doi.org/10.1037/apl0000365

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*, 445-494. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology, 4*, 653-663. https://doi.org/10.1037/0021-9010.68.4.653

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117-143. https://doi.org/10.1037/pspp0000096

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500-517. https://doi.org/10.1037/0021-9010.88.3.500

Wang, Q., & Bowling, N. A. (2016). A comparison of general and work-specific personality measures as predictors of organizational citizenship behavior. *International Journal of Selection and Assessment, 24*, 172-188. https://doi.org/10.1111/ijsa.12139

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

Wilmot, M. P., & Ones, D. S. (2019). A century of research on conscientiousness at work. *Proceedings of the National Academy of Sciences, 116*, 23004-23010. https://doi.org/10.1073/pnas.1908430116

Woo, S. E., Jin, J., & LeBreton, J. M. (2015). Specificity matters: Criterion-related validity of contextualized and facet measures of conscientiousness in predicting college student performance. *Journal of Personality Assessment, 97*, 301-309. https://doi.org/10.1080/00223891.2014.1002134

Zayas, V., Whitsett, D., Lee, J. J. Y., Wilson, N., & Shoda, Y. (2008). From situation assessment to personality: Building a social-cognitive model of a person. In G. Boyle, G. Matthews, & D. Saklofske (Eds.), *Handbook of personality theory and testing* (pp. 208-217). Sage.

Notes

1 Although there are other methods of assessing reliability (e.g., test-retest), test developers do not always have the resources to use alternative approaches.

2 Only a portion of those who consented were subsequently rated by a supervisor or coworker.

3 Previous research suggests that MTurk can be an adequate data source (Paolacci & Chandler, 2014) and that novices can be used to rate SJT items (Motowidlo & Beier, 2010). We verified the ratings provided by MTurk workers for the 12 SJT items in the current study against ratings provided by 18 I/O psychology graduate students and found them to be very similar (r = .98). SJT scores calculated with the point values obtained from the MTurk raters and SJT scores calculated using the point values obtained from the graduate student raters showed essentially the same correlations with the other variables in the current study.

4 Our initial data collection did not include a fixed SJT form where all participants answered the same set of questions.

5 Reversing ratings for low achievement response options addresses the issue that within SJT item, raw likelihood ratings for low achievement options correlate negatively with likelihood ratings for high achievement options.

6 We split one of the items from the original scale ("Set high standards for myself and others") into two because it appeared to be double-barreled.

7 To correct for indirect range restriction, we used Salgado and Táuriz's (2014) conscientiousness range restriction value of .88 and Thorndike's case 3 correction formula (see Sacket & Yang, 2000).

8 Results of analyses were substantively the same when task performance was not dichotomized and was used in OLS regressions instead.

Cite this article as: Golubovich, J., Lake, C. J., Anguiano-Carrasco, C., & Seybert, J. (2020). Measuring achievement striving via a situational judgment test: The value of additional context. *Journal of Work and Organizational Psychology*, 36(2), 157-168. https://doi.org/10.5093/jwop2020a15

Funding: This research was supported by the Educational Testing Service and Kansas State University.

Author notes

Correspondence: jgolubovich@gmail.com (J. Golubovich).

Conflict of interest declaration

Conflict of Interest

The authors of this article declare no conflict of interest.