SITUATIONAL JUDGMENT TESTS AND PSYCHOLOGICALLY ACTIVE CHARACTERISTICS OF SITUATIONS: A DIMENSIONAL APPROACH TO ANALYZING SITUATIONAL JUDGMENT TEST CONTENT AND ITS PSYCHOMETRIC IMPLICATIONS

By

Matthew C. Reeder

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Psychology – Doctor of Philosophy

2013

ABSTRACT

SITUATIONAL JUDGMENT TESTS AND PSYCHOLOGICALLY ACTIVE CHARACTERISTICS OF SITUATIONS: A DIMENSIONAL APPROACH TO ANALYZING SITUATIONAL JUDGMENT TEST CONTENT AND ITS PSYCHOMETRIC IMPLICATIONS

By

Matthew C. Reeder

At a very basic level, a situational judgment test (SJT) is a series of situations and associated behaviors relevant to each situation. Although a sizable body of research relevant to SJTs has accrued, little is known about how properties of situations and behaviors, as fundamental units in SJT design, are related to properties of SJTs in terms of the information provided by scores at the item-level. In this study, theory and empirical research relevant to situations, interactionism, and trait activation provide a foundation for the argument that situational and behavioral characteristics would explain item-level variability in relationships with external variables, namely other individual difference characteristics and criterion-related validities. Ninety items from three SJTs were coded with regard to item stem situational cues and factor-five model personality trait expression associated with the response options. Mixed support was found for the study's assertions in analyses pertaining to two types of SJT scores (stem-level scores and response option-level scores). Response option personality trait expression was significantly related to response option correlations with like personality characteristics. Further, models predicting item-level correlations for both stem scores and response option-level scores explained a respectable proportion of between-item variability in

correlations with external variables. Finally, there was strong evidence that the effect of response option trait expression clustered around or varied significantly across the item stems within which the response options were nested. However, results were inconsistent with regard to the effect of situational characteristics on stem score-level correlations with external variables. Additionally, results pertaining to interactions between situational characteristics and behavioral characteristics in predicting response option-level correlations with external variables were mixed. The dissertation concludes with a discussion of the study's implications for trait activation and the design of SJTs and other similar measurement procedures that rely on the sampling of situational content (e.g., work samples, assessment centers).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

As a measurement procedure that has become increasingly utilized in organizational and educational contexts, situational judgment tests (SJTs) are a useful tool for attempting to understand what respondents think they would or should do when confronted with the demands and constraints that define the environment in question (e.g., the workplace, the classroom). SJTs fall within a class of selection procedures that includes interviews, work samples, and assessment centers (ACs). Collectively, procedures within this class are frequently referred to as simulations (Motowidlo, Dunnete, & Carter, 1990) or situational tests (Weekley & Jones, 1997; 1999). The commonality among these procedures is the use of stimuli that are sampled and designed to imitate domain-representative situations, with the intent of eliciting responses that can be interpreted as indicators of how individuals would behave within the situation (Motowidlo et al., 1990). Similar to other simulation methods, the use of SJTs is often motivated from a philosophy of behavioral sampling, in accord with arguments that measures of behavioral samples should be more predictive of performance and other criteria than are measures of trait-like predispositions (e.g., Goodenough, 1949; Wernimont & Campbell, 1968).

SJTs have been employed in selection and admissions contexts for almost 100 years (for historical reviews, see McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006). However, systematic streams of research on SJTs largely appeared only within the past 15 to 20 years, following the publication of seminal studies by Motowidlo and colleagues (Motowidlo et al., 1990; Motowidlo & Tippins, 1993) and others (e.g., Dalessio, 1994). Interest concerning the psychometric characteristics of scores from SJTs subsequently grew. For instance, researchers began investigating relationships between SJT scores and other individual characteristics in the domains of personality, cognitive

ability, knowledge, and experience (MacKenzie, Ployhart, Weekley, & Ehler, 2010; McDaniel & Nguyen, 2001; Schmitt & Chan, 2006; Weekley & Ployhart, 2005), relationships with criteria in both organizational and educational settings (e.g., McDaniel, Hartman, Whetzel, & Grubb, 2007; McDaniel et al., 2001; Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004), and subgroup differences in test scores (e.g., Chan & Schmitt, 1997; Motowildo & Tippins, 1993).

In addition to investigations of the psychometric characteristics of SJTs, researchers have addressed design and development issues associated with SJTs. Examples include research on the comparability of media available for the administration of SJT content (e.g., video-based administration, computer-based administration; Chan & Schmitt, 1997; Dalessio, 1994; Weekley & Jones, 1997), the implications associated with the use of different response instructions (e.g., "should-do" versus "would-do" instructions) in terms of validity and subgroup differences (McDaniel et al., 2007; Ployhart & Ehrhart, 2003), and the use of various keying strategies (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006; Weekley & Jones, 1997; 1999). Other practical matters that have been examined include comparisons of SJT score properties in applicant versus incumbent samples (MacKenzie et al., 2010), the influence of response distortion on the validity of SJT scores (Nguyen, Biderman, & McDaniel, 2005; Peeters & Lievens, 2005), the effects of retesting on SJT scores (Lievens, Buyse, & Sackett, 2005), and parallel test form development (Clause, Mullins, Nee, Pulakos, & Schmitt, 1998; Lievens & Sackett, 2007; Oswald, Friede, Schmitt, Kim, & Ramsay, 2005).

Given that the majority of the systematic research on SJTs has been conducted only within the past two decades, the brief sketch of the literature presented above indicates the substantial progress made in knowledge concerning SJTs as measurement procedures, properties of scores derived from SJTs, and design and applied issues associated with the implementation

and operational use of SJTs. In spite of these advances, however, there has been little empirical research conducted on what is argued here as a critical design feature of SJTs: the psychological characteristics associated with SJT item content, including item stems and response options.

This gap is somewhat surprising, as numerous researchers have commented on how the lack of an understanding of characteristics of SJT items impedes further conceptual understanding of SJTs. For instance, Schmitt and Chan (2006) commented on the relative paucity of research concerning how SJT content influences test scores. At the same time, however, researchers have acknowledged that characteristics of SJT content likely represent an important source of variation in scores (Bledow & Frese, 2009). Similar observations have been made concerning other forms of situational testing such as assessment centers (e.g., Neidig & Neidig, 1984; Sackett & Dreher, 1984). Thus, one may surmise that SJT content is likely related to properties of SJT scores, but that there exists a relative lack of research that explicitly addresses this matter. This matter represents the chief concern of the present dissertation. In order to facilitate understanding for why features of SJT content might be associated with psychometric properties of SJT scores, theory from personality psychology on interactionism and behavioral consistency is incorporated as a conceptual foundation.

Historically, attributes in the domain of personality have often been treated as global aggregates, dispositions, or propensities that are relatively de-contextualized. In this view, variability in an individual's behavior across different situations is conceptualized as specificity or measurement error to be aggregated over to the benefit of reliability. However, more recent approaches to personality that emphasize psychologically-relevant features of situations (e.g., Mischel & Shoda, 1995; 1998) characterize personality in terms of relatively stable behavioral signatures. These signatures represent patterns of variability in behavior across specific situations

and can be viewed as *if…then* relations (Mischel, 1994; Shoda, Mischel, & Wright; 1994). The *if* component of the behavioral signature corresponds to the psychological situation as it is presented to and perceived by the individual; the *then* component represents the behavior chosen by the individual in response to the situation (Smith, Shoda, Cumming, & Smoll, 2009).

SJTs and other similar simulation-based methods are often viewed as advantageous in part because they allow the observation of samples of behavior in response to the diverse array of situational demands that characterize organizational and educational reality (Mumford, Campion, & Morgeson, 2006). The notion of an *if…then…* conditional behavioral signature corresponds in many respects to the format of a typical SJT item. The item stem of the SJT, which describes the situation or incident as it is presented to the respondent, reflects the *if* in a behavioral signature. The behavioral response options for each SJT item, on the other hand, correspond to a series of possible *then*s from which a respondent is instructed to select in light of the situation or incident as described in the item stem. In a sense, then, SJTs can then be viewed as measures of behavioral signatures.

An important implication of research and theory on behavioral signatures and behavioral consistency is that variability in response across situational content may not simply reflect error in measurement as might be suggested by certain dispositional- or trait-based interpretations of personality. Rather, if the psychological characteristics associated with two situations differ appreciably, and if these differences cause variability in behavioral response, then there is no reason to view the two situations as parallel indicators of a given construct. In other words, if behavioral response varies in a systematic manner conditional on changes in the characteristics of situations that confront respondents, then this lack of consistency may not reflect error and

should not simply be treated as such. Rather, it may reflect the meaningful influence of the situations on characteristics of respondents' responses.

Viewing SJTs in this light, one might predict a greater degree of response consistency across responses (the *then*s in the behavioral signature approach) if the situational content represented in the item stems (the *if*s in the behavioral signature approach) is similar or equivalent in terms of meaningful psychological characteristics. If the psychological characteristics associated with the situations or incidents described in stem content are not similar or equivalent, however, there is no reason to expect consistency in response. Rather, response inconsistency would result in psychometric characteristics such as relatively low internal-consistency estimates and complex or theoretically un-interpretable factor structures; as will be elaborated upon shortly, such outcomes are not infrequent in SJT research. This perspective might also suggest that other psychometric features of items (e.g., convergent or discriminant validities between the items and other individual difference characteristics, criterion-related validities of the individual items) would be expected to vary across items, as well. If psychometric characteristics do vary across items systematically as a function of the psychological features associated with the test content, this information might be leveraged in a way to approach SJT design from a standardized, evidence-based perspective.

To recapitulate, the proposed study incorporates ideas from research in personality and social psychology on behavioral consistency, interactionism, and the psychological characteristics of situations with research on the design and validity of SJTs. Research relevant to characteristics of content in personality inventories (e.g., Werner & Pervin, 1986; Rauthmann, 2011; Rauthmann & Denissen, 2011; Zickar & Ury, 2002) and cognitive ability tests (e.g., Kobrin, Kim, & Sackett, 2012) has been conducted; however, there has been little in the way of

similar research for SJTs in the published literature. Furthermore, there is little empirical evidence or theory concerning psychologically active characteristics specific to SJT item content. There is no agreed-upon taxonomy of situations in the organizational or personality/social psychological literatures, although potential frameworks are emerging (e.g., Tett & Burnett, 2003). Therefore, as elaborated upon shortly, a hybrid deductive-inductive approach was applied toward the analysis of situational characteristics relevant to SJT item content.

In short, the basic idea underlying the proposed study is to: (a) scale SJT item content in terms of psychological features derived from existing research and theory concerning situational characteristics and behavioral consistency, and (b) link those characteristics to psychometric properties of SJT response data, namely (1) item-level correlations with individual difference characteristics (personality, knowledge, experience), and; (2) item-level criterion-related validity. Findings will be used to argue that such psychologically active features embedded in SJT item content, and potentially other complex, multidimensional measurement systems (e.g., assessment centers), can be used in a theoretically-motivated manner in the systematic development and application of these selection procedure in high-stakes, operational settings.

The literature review is structured as follows. First, research in three areas is reviewed: (1) measurement research on SJTs; (2) personality and social psychological research on behavioral consistency and situational characteristics, and (3) organizational research relevant to the study of situations in the workplace. Following this review, a description is provided of the research to be conducted per the present research, including the study hypotheses that will be addressed. Finally, the Method section elaborates upon the development of an inventory designed to measure the psychological features of content in simulation-based measurement procedures. As elaborated upon in the forthcoming sections, development of this inventory draws

upon theoretical frameworks in both the personality-social psychological literatures on situational characteristics and research in organizational settings concerning features associated with contextual characteristics. Also discussed in the Method section are the sampling of respondents and SJT content that will be the target of study, as well as the analytic procedures used to test the proposed hypotheses. Following the Method section is the presentation and discussion of the study results. The dissertation concludes with implications, a review of strengths and weakness, and suggestions for future research.

LITERATURE REVIEW

This review elaborates upon the ideas presented above concerning the conceptual rationale for the proposed study, to summarize what is known concerning the relevance of psychological characteristics of situations for behavioral consistency, and to apply these ideas to understanding the validity of SJT scores. First, a brief review of SJTs as a measurement method is provided. A selective emphasis is placed on research associated with design and development features, criterion-related validity, and construct validity. This review concludes with two observations: (1) scores from SJTs are useful from the standpoint of outcome prediction in applied settings; (2) SJT scores are inherently multidimensional, the sources of this multidimensionality are not well understood, and that this lack of understanding may impede further developments in SJT understanding and application. One avenue toward understanding multidimensionality in SJT scores lies in understanding the psychological characteristics inherent in SJT item content.

Second, research from personality and social psychology on the psychological characteristics of situations is discussed as it pertains to the consistency of behavior across situations. Emphasis is placed on studies that examine behavioral consistency from the standpoint of how situations and situational features contribute to consistency (or inconsistency) in behavior. This review provides a perspective from which to interpret findings concerning the multidimensionality of SJT scores, and argues that consistency in responses to SJTs should not simply be assumed a priori. Indeed, research over the past 40+ years on behavioral consistency and interactionism foreshadowed some of the recent empirical findings concerning SJT multidimensionality.

Finally, relevant theory and research concerning situations and behavioral consistency in organizational psychology is reviewed. This discussion focuses on three broad topics, some of which are interrelated, that have received attention and that are relevant for the present discussion. These include: (1) the effects of situational strength and situational constraints as influences on behavior and the prediction of behavior in organizational settings; (2) the notion of trait activation, which connects situational strength with the notion of trait relevance to make predictions regarding when individual differences in personality are manifest in workplace behavior, and; (3) the application of theory regarding behavioral consistency and the psychological nature of situations to research on assessment centers.

Situational Judgment Tests: A Review

SJT items have two basic components. First, the respondent is presented with an *item stem*, a hypothetical situation or incident similar to that which might occur in a workplace or educational environment. SJT item stems frequently contain content that reflects situations that have actually occurred within the setting for which the SJT has been designed. However, item stems may also contain situational content that is isomorphic to the types of situations that could feasibly occur (Schmitt & Chan, 2006). Second, the respondent is provided a series of behavioral *response options*, generally three to twelve in number, each of which represents a course of action that one could select in light of the situation (McDaniel & Nguyen, 2001; Weekley, Ployhart, & Holtz, 2006). Oftentimes, there is no ostensibly "correct" or "incorrect" response to each situation in an SJT. Rather, multiple response options might be effective, perhaps to varying degrees, depending on the demands and constraints that characterize the incident in question

(Bergman et al., 2006). Instructions prompt the respondent to consider the situation, and to make a judgment among the available response options.

SJT Development and Application

SJT development entails a number of considerations. Broadly speaking, such considerations pertain to the content of item stems and response options, the response instructions that are provided to respondents, the approach used for scaling responses, and the means by which responses are keyed (Weekley et al., 2006).

Two approaches are frequently used for item stem and response option development: the critical incident technique (Anderson & Wilson, 1997; Flanagan, 1954; Smith & Kendall, 1963) and theory-based content development (Weekley et al., 2006). SJT development based on critical incidents is inductive in nature. The process begins by observing what exists within the context of interest and proceeds by applying these observations to develop theory concerning the relevant situational and behavioral domains that drive test development. Specifically, descriptions of critical incidents are collected from subject matter experts (SMEs; e.g., incumbents, supervisors), reflecting instances of unusually effective or ineffective performance. These critical incident descriptions include information concerning the events that transpired, the actions that were taken in response to the events, and the consequences associated with those actions (Anderson & Wilson, 1997). The test developer reviews the incident descriptions with the intent of identifying a series of situations that will serve as item stems (McDaniel & Nguyen, 2001). Frequently, a second group of SMEs is then tasked with providing responses to each incident, although sometimes the test developer will develop the response options (e.g., Weekley & Ployhart, 2005).

Theory-based SJT development, on the other hand, is generally deductive in nature. Rather than basing test construction off of what is observed in a specific context, theory-based development concerns the application of a conceptual framework regarding a construct or domain of interest to guide measure development. For instance, a test developer may construct an SJT designed to measure individual differences in characteristics associated with, for example, customer service, teamwork, conflict management, or honesty and integrity. Unlike tests developed using the critical incident method, there may be no explicit attempt in theory-based development to sample situational content in order to ensure that certain basic considerations are attended to (e.g., contextualizing the incidents, using only incidents that are feasible in the general context in question). In the case of SJTs that are developed around a specific theory or construct, the test designer constructs a pool of situations that are created to elicit an underlying theoretical construct or dimension, as well as response options for each item stem that serve as indicators of the construct of interest (Weekley et al., 2006).

Response instructions and the scaling of responses are related considerations in SJT development. Two approaches used to scale SJT responses include multiple- or forced-choice methods (e.g., where respondents are prompted to choose a subset of the available responses) and Likert-type methods (e.g., where each response option associated with a given stem is rated on a seven-point scale; Chan & Schmitt, 1997). Although a variety of options exist with regard to SJT response instructions (McDaniel & Nguyen, 2001), two broad classes of instructions are frequently distinguished: knowledge or "should-do" instructions and behavioral tendency or personality instructions (McDaniel et al., 2007). Examples of knowledge instructions include those that prompt the respondent to indicate what she or he should do in response to each situation or what she or he thinks is the best or most effective response. Respondents may also be

asked to choose two or more options (e.g., a best and a worst response, a best and a second best response), or to rate all of the individual response options with respect to their effectiveness in light of the incident (Ployhart & Ehrhart, 2003). Examples of behavioral tendency or "would-do" instructions include prompting the respondent to indicate what she or he would do or would have done in light of each situation. Similarly, respondents may be asked to indicate what they would be most and least likely to do. The choice of response instructions has been found to influence response distributions, reliability, and validity (e.g., Ployhart & Ehrhart, 2003), indicating that decisions concerning response instructions are not immaterial.

Finally, a number of strategies exist for keying SJTs. Bergman and colleagues (2006) discuss four classes of keying methods, including empirical keying, theoretical keying, expert-based keying, and hybrid keying. Empirical keys are developed on the basis of statistical evidence concerning relationships between responses and some criterion of interest. Theoretical keying entails the scoring of items in accordance with the construct that the SJT in question was designed to assess, with behaviors that reflect the presence of the construct being scored positively and behaviors that reflect the absence or the opposing end of the theoretical continuum being scored negatively. Expert-based keying strategies employ SMEs with regard to the job or the constructs in question to determine the effectiveness of the response options given the incidents presented in the item stems. Finally, hybrid keys entail strategies that combine other keying methods, for instance, use of a theoretical and empirical keying strategy in tandem.

As mentioned above, SJTs that are designed where each item stem is associated with a series of response options are generally scored in one of two ways, depending on item format. Some SJTs utilize a multiple-choice or forced-choice format, where the respondent is prompted to choose specific response options as corresponding to the best and worst options, the options

that one would be most and least likely to carry out, and so forth. The use of a multiple-choice format often entails the use of scoring protocols that result in *stem scores*, where the respondent receives a score for each stem or situation.

As an example of this approach used with best/worst response instructions, a scoring key might first be developed by administering the item stems and response options to a group of SMEs and instructing them rate each response option in terms of effectiveness. Thus, for each item stem, some responses are high in effectiveness, some are neutral in effectiveness, and some are low in effectiveness. When administered operationally, a respondent would get a 1 for selecting a highly effective option as the best option, a 0 for selecting a neutral option as the best option, and -1 for selecting an ineffective option as the best option. Similarly, a respondent would get a 1 for selecting a highly ineffective option as the worst option, a 0 for selecting a neutral option as the worst option, and a -1 for selecting an effective option as the worst option. Scores are then summed across response options for each stem, resulting in stem scores that may vary from -2 (for an individual who chose the most ineffective option as the best and the most effective option as the worst) to +2 (for an individual who chose the most effective option as the best and the most ineffective option as the worst). For an early example of this approach, see Motowidlo et al. (1990).

Other SJTs utilize a rating-scale format, where the respondent is prompted to rate all response options according to effectiveness or some other criterion. Whereas use of the multiple-choice or forced-choice format results in stem scores, use of the rating-scale format generally does not. Again, the rating-scale format generally entails prompting respondents to rate all response options for each item stem in terms of some criterion such as effectiveness. As is done with other measurement procedures in selection practice that rely upon rating-scale item formats

(e.g., self-report personality inventories), scores are aggregated across the ratings for the response options, often across the entire test (i.e., without regard to the specific item stems). Thus, instead of resulting in scores specific to each stem that are aggregated to yield a test score, the ratings applied to the individual response options are scored and then aggregated across all stems to yield a test score. This distinction between stem scores and response-option scores is relevant to discussion of the study hypotheses presented shortly.

SJTs have been considered as applicable in two areas of operational use. The vast majority of research addresses the development, validation, and application of SJTs for use in high-stakes testing situations where individuals will be selected on the basis of test scores for positions within organizational or educational settings (e.g., Motowidlo et al., 1990; Motowidlo & Tippins, 1993; Oswald et al., 2005). However, there has also been some consideration of the use of SJTs in training contexts (e.g., Fritzsche, Stagle, Salas, & Burke, 2005). For instance, Hauenstein, Findlay, and McDonald (2010) discussed the application of SJTs within the context of training equal opportunity advisors in the United States Armed Forces. As Hauenstein and colleagues noted, SJTs have a number of potential uses within training contexts, including serving as proxies for transfer of training, providing a diagnostic tool for identifying gaps in training efforts, and facilitating development. Moreover, the mental simulation of situations or events (either reconstruction of past events or rehearsals of likely future events) facilitates the translation of cognition into behavior, assists with planning, and fosters motivation (e.g., Taylor & Schneider, 1989). Thus, although the vast majority of research on SJTs has focused on application within selection and admission contexts, the application of SJTs may also be useful in training contexts.

The Criterion-Related and Construct Validity of SJTs

For SJTs applied operationally in selection and admissions contexts, additional considerations beyond design and development concern criterion-related validity, convergent validity, and discriminant validity. Prior to systematic public research on SJTs, the predictive validity of SJTs was forecasted almost half a century ago by researchers studying behavioral consistency and interactionism. For instance, Endler and Hunt (1966) argued that "…the validity of predictions of personal behavior should be substantially improved by asking the individuals concerned to report the trait-indicating responses of interest in the specific situations…concerned" (p. 343). Wallace (1966) similarly suggested that prediction should improve to the extent that the situation within which the respondent enacts behavior approximates characteristics of the "predictive situation." Similar arguments have been put forth in the organizational literature concerning selection procedures that utilize the observation of behavior within situations designed to be sampled from the job or position in question (e.g., Asher & Sciarrino, 1974; Wernimont & Campbell, 1968).The implication of arguments such as those put forth by Endler and Hunt (1966), Wallace (1966), and others is clear: prediction of behavior benefits by knowing not only how respondents indicate they should or would behave, but also by knowing something about the specific situational characteristics that define the circumstances surrounding decisions regarding behavior.

Meta-analytic summaries provide evidence for the criterion-related validity of SJTs. McDaniel and colleagues (2001) reported an uncorrected mean validity of .26 (ρ = .34) based on 102 estimates. More recently, McDaniel et al. (2007) reported a slightly lower overall validity of .20 (ρ = .26), based on 118 estimates. Christian, Edwards, & Bradley (2010) categorized SJTs into construct domains (e.g., knowledge and skills, applied social skills, personality) and, within each domain, specific construct categories (e.g., interpersonal skills, teamwork skills, and

leadership fell within applied social skills). Although the number of studies associated with the specific categories was sometimes low (*k* ranging from 4 through 51), criterion-related validities for all categories were non-zero and demonstrated validity generalization.

The above-cited meta-analytic evidence illustrates the potential usefulness of SJT scores for predicting performance. However, also of interest to the present study are the credibility intervals associated with these validity estimates. The interpretation of credibility intervals addresses the existence of potential moderators of the relationships under investigation (Whitener, 1990). The moderator of interest for this study pertains to the psychological features of SJT content. Whereas the focus of the present study is on item-level psychological characteristics, it is also reasonable to assume that item-level differences aggregate to test-level differences (although this is, to some degree, conjecture; the question is subject to empirical examination). If psychological characteristics inherent in SJT test content vary across tests and if such characteristics influence validity, one expectation would be meta-analytic credibility intervals about the point estimates that are wide enough to indicate that validities vary after controlling for study artifacts.

Although SJT scores do evince validity generalization in the sense of credibility intervals that do not overlap with zero, the intervals are not restrictively narrow. With respect to relationships with job performance, the width of the reported 80% credibility intervals in Christian and colleagues' (2010) meta-analysis ranged from .17 (80% CV of .35–.52 for personality composites) to .29 (80% CV of .24–.53 for teamwork skills). McDaniel et al. (2007) reported 80% credibility intervals of .13 to .39 around the overall estimate reported in their meta-analysis (see Table 3, p. 72). Interestingly, when test content was held constant across studies by examining only primary studies where the same test was administered, the width of the

credibility intervals approached zero (McDaniel et al., 2007; Table 5, p. 75). Thus, holding test content, and hence the psychological characteristics of SJT content, constant across studies reduces variability in SJT validity estimates. Fixing test content in this manner is a stringent method of examining the effects of the psychological characteristics of SJT content on criterion-related validity at the test level. Clustering SJTs into global composite categories such as teamwork skills or leadership skills, as was done by Christian et al. (2010), is a less stringent approach, as there may still exist within-category test-level heterogeneity with regard to characteristics represented in the stems or response options.

In addition to the research findings that have accumulated on the criterion-related validity of SJTs, a related body of research has developed with regard to SJT convergent and discriminant validity. Recent reviews conclude that little is known about the construct validity of SJT scores (Christian et al., 2010) or about developing SJTs that yield scores that reflect unidimensional construct domains (De Meijer, Born, van Zielst, & van der Molen, 2010; MacKenzie et al., 2010). Concurrently, researchers have frequently commented upon the apparent multidimensionality or construct heterogeneity of SJTs. For instance, Chan and Schmitt (1997; 2002; see also McDaniel & Nguyen, 2001) suggested that situational judgment items draw upon multiple domains of individual difference characteristics (e.g., personality, ability). Given that performance in organizational or educational settings is associated with various domains of person characteristics, it thus makes sense that SJTs may not provide "clean" measurement of specific constructs (Christian et al., 2010). However, although scores from SJTs are likely to be multidimensional, the conditions responsible for this multidimensionality are not well understood.

These conceptual arguments have been addressed empirically through various construct validation strategies, namely those associated with internal structure, correlations with external variables, and the examination of SJT content. With respect to internal structure, relatively few published studies on SJTs actually report attempts to examine the factor structure of SJTs. Exceptions do exist (e.g., Chan & Schmitt, 1997; 2002; Clause et al., 1998; Oswald et al., 2004; 2005), with several studies reporting that extraction of a single factor yields a solution that accounts for a very small percentage of shared variance, whereas extraction of multiple factors does not frequently yield theoretically or rationally interpretable solutions. Thus, simpler solutions appear to result in poor model fit in terms of the model's ability to account for variability in response data, whereas more complex solutions do little to elucidate rationally or theoretically interpretable dimensions that explain response data.

In a related vein, many studies report internal-consistency reliability estimates that would generally be considered low for tests to be used for making personnel decisions (e.g., Bledow & Frese, 2009; Chan & Schmitt, 1997; Clause et al., 1999; Motowidlo & Beier, 2010; Mumford et al., 2008; Weekley & Jones, 1997). Findings of low reliability have led researchers to suggest that methods of reliability estimation based on internal-consistency methods such as Cronbach's α are inappropriate for use with SJT response data (McDaniel et al., 2007). High estimates of internal consistency (i.e., those approaching unity) are insufficient to conclude unidimensionality; however, low estimated internal consistency may reflect multidimensionality, particularly in situations where the number of items under consideration is not prohibitively small (Cortina, 1993). As a consequence of the aforementioned multidimensionality, low observed internal-consistency reliability, or both, SJT developers have frequently resorted to reporting overall composite scores, even when the test was explicitly developed with multiple

theoretical constructs or domains in mind (e.g., Motowidlo et al., 1990; Motowidlo & Tippins, 1993; Oswald et al., 2004).

Concerning relationships between SJT scores and external correlates, researchers have primarily focused on various domains of individual difference attributes, including personality, cognitive ability, experience, and knowledge. With reference to personality, McDaniel and colleagues (2007) found that correlations between SJT scores and Agreeableness, Conscientiousness, and Emotional Stability (mean ρ of .25, .27, and .22, respectively) were somewhat stronger than those associated with Extraversion and Openness to Experience (mean ρ of .14 and .13, respectively). However, only a small percentage of the between-study variance in the correlations for most of these relationships was accounted for by statistical artifacts (for Openness, 66% of the variance was attributable to artifacts; for the remaining five-factor model characteristics, 25% or less of the variability in estimates was attributable to artifacts).Thus, whereas the mean estimates indicate some degree of convergence with five-factor model (FFM) personality characteristics, the magnitude of the relationships varies across studies. As was discussed above concerning variance in reported relationships between SJT scores and job performance, McDaniel et al. (2007; see Table 5, p. 75) found that holding test content constant across studies drastically reduced the percentage of variance in estimates not due to artifacts for relationships between FFM characteristics and SJT scores (for all FFM characteristics, the majority of the variance in estimates was attributable to statistical artifacts). Although not definitive, this again suggests that fixing the psychological features of the situations in SJT test content by holding content constant may be useful for understanding psychometric features of SJT scores, in this case, correlations with FFM traits.

Researchers have argued that SJT scores should be related to cognitive ability (e.g., Weekley & Jones, 1999). Accordingly, empirical examinations of the SJT-cognitive ability relationship have been conducted (e.g., Chan & Schmitt, 2002; Clevenger et al., 2001; Weekley & Jones, 1997; 1999), with varying results. McDaniel and colleagues (2001) report a corrected estimate of the SJT-cognitive ability relationship of .39 (k = 80 studies), somewhat higher than the more recent estimate of .32 from McDaniel et al. (2007) based on 95 studies. Further, McDaniel and colleagues (2007) found some evidence that the relationship between SJT scores and cognitive ability varies depending upon response instructions, with higher estimates for knowledge instructions (.35) than for behavioral tendency instructions (.19). Again, however, only a small amount of the variability in the estimates across studies was due to statistical artifacts; the percentage of variance in estimates attributable to artifacts again rose when SJT content was fixed across studies.

Because SJTs include item stems associated with domain-relevant situational content, relationships with experience and knowledge might be expected (Clevenger et al., 2001). Estimates of validities between SJT scores and experience have generally been fairly small (e.g., between .05 and .25), although results are somewhat varied (e.g., Chan & Schmitt, 2002; Weekley & Jones, 1997; 1999). In support of this, McDaniel and Nguyen (2001) reported a mean estimated correlation of .05 between SJT scores and experience across 18 studies, although less than 20% of the variance in estimates across studies was accounted for by sampling error (see Table 1, p. 108).

In addition to considering construct validity evidence associated with internal structure and correlations with external variables, several initial attempts have been made to empirically examine characteristics of SJT content that might be relevant to scores. At the broadest level of

such attempts is Christian and colleagues' (2010) classification SJTs in terms of constructs assessed on the basis of test content. Christian et al. adopted Huffcut, Conway, Roth, and Stone's (2001) taxonomy, originally developed for interviews, to yield a taxonomy based on four broad domains of characteristics: knowledge and skills, applied social skills, personality, and heterogeneous composites. Each of these domains, with the exception of heterogeneous composites, encompassed multiple characteristics, including job knowledge and skills (knowledge and skills), interpersonal skills, teamwork skills, and leadership (applied social skills), and personality composites and conscientiousness (personality).

As opposed to Christian and colleagues' (2010) categorization of tests into broad categories on the basis of test-level judgments, a number of researchers have focused on the content contained in item stems or response options. For instance, ratings associated with the behavioral response options contained in SJTs have been collected in several studies, generally with regard to either the perceived effectiveness of the response option or the extent to which the response option is perceived to be indicative of a given characteristic such as agreeableness or conscientiousness (e.g., Bledow & Frese, 2009; De Meijer et al., 2010; Kell, Rittmayer, Crook, & Motowidlo, 2010; Motowidlo & Beier, 2010).

Other researchers have attempted to categorize or rate features of item stems. For instance, two of the authors in the Motowidlo et al. (1990) study rated item stems with respect to the extent to which they reflected interpersonal or problem-solving "elements." Similarly, Kell et al. (2010) examined the extent to which various item stems provided opportunities to demonstrate helping behaviors or task-related behaviors, yielding some evidence that these two aspects were moderately to strongly related to one another in a negative fashion (i.e., situations that provided more opportunity for task-relevant behavior provided less opportunity for helping

behavior and vice versa). As argued later, two related limitations of rating or categorizing item stems or incidents using the task/problem-solving versus social/interpersonal/helping distinction is that it (a) ignores more specific situational features or cues that have been discussed in the personality, social, and organizational literatures, and (b) is arguably too gross or broad for purposes such as generation or development of content at the individual item level.

In summary, research on the criterion-related validity and construct validity of SJT scores at the test-level demonstrates that (1) SJT scores are predictive of performance, (2) SJT scores have a large number of individual difference correlates that have been examined empirically, and (3) meta-analytic estimates of criterion-related validity and construct validity frequently exhibit considerably wide credibility intervals, suggesting that estimates of validity vary systematically across studies. Each of these points is relevant, however, to scores at the level of the individual test. One of the primary goals of the present study is to delineate situational characteristics in SJT content relevant for influencing the psychometric properties of SJTs at the item level. First, however, it is useful to review prior research from related fields (e.g., personality) concerning the psychological features of situations and their relevance for behavioral consistency. Such a review serves two purposes. First, it highlights some of the situational characteristics that have emerged in prior research or that have been theorized as being important influences on behavior. Second, it relates behavioral consistency to psychological features of situations, a useful foundation from which to begin understanding why features of situational content in SJTs might influence the psychometric properties of SJT scores.

Psychological Characteristics of Situations and Behavioral Consistency

Social scientists, and psychologists in particular, have long theorized about the functional significance of situational forces on behavior (for reviews, see Ekehammar, 1974; Pervin, 1978).

The conceptual contributions that have been made reflect the array of theoretical orientations (e.g., Gestalt psychology, behaviorism, social-cognitive, interactionism) that have been applied by researchers who have grappled with the concept of situation. Despite the long history of situational theorizing, researchers for over 50 years have repeatedly bemoaned the lack of a coherent, agreed upon taxonomy of situational characteristics (e.g., Jessor, 1956; Magnusson, 1971; Monson, Hesley, & Chernick, 1982; Pervin, 1976; Reis, 2008; Sherman, Nave, & Funder, 2010; Ten Berge & De Raad, 1999) and the related lack of a methodology for assessing characteristics of situations (e.g., Frederiksen, 1972; Magnusson, 1971; Sherman et al., 2010). Organizational psychologists and management scholars have also noted this state of affairs (e.g., Beaty, Cleveland, & Murphy, 2001; Kell et al., 2010; Weekley et al., 2006). Although comprehensive classification or taxonometric structures of situational characteristics have not been developed, a large number of specific situational characteristics have been studied. Prior to discussing these characteristics and their relevance for understanding behavioral consistency, a useful first step is to understand exactly what is meant by the term "situation."

Defining Situations

Situations are comprised of certain basic components. These components include some consideration of who is involved in the situation (permitting for the possibility that the individual is alone), the nature of the action or activities that are transpiring, and where the actions or activities are taking place (Pervin, 1978). Thus, situations contain elements associated with actors, behaviors, and settings. Actors can be the self, others, or both. Behaviors are often observable, overt actions, although cognitions, affects, and other non-observable phenomena are also often relevant (e.g., the actor may be introspecting, baking, or mourning). The setting can be the home, the corner delicatessen, or the conference room. Collectively, the elements that

comprise the situation are perceived oh as contributing to a "wholeness" or gestalt quality associated with the situation. Therefore, modification or alteration of any one component affects the perception of the situation as a whole (Pervin, 1978) and, hence, the influence of the perceived situation on behavior.

Situations are distinct from environments, although the boundary between these concepts is not often well-defined. Environments can be thought of as general, persistent, or relatively stable contexts or settings within which action occurs (Endler, 1981). Researchers interested in the study of human environments have frequently focused on objective characteristics, although environmental characteristics are not restricted solely to those features that are objective or nominal in nature (Moos, 1973). In the organizational sciences, topics associated with the study of environment include organizational culture and climate (e.g., James & Jones, 1974; O'Reilly, Chatman, & Caldwell, 1991; Schneider & Reichers, 1983). Relative to environments, situations are conceptualized as being more transient and temporary in nature, representing episodes or events comprised of specific stimuli that serve as the target of attention and response (Pervin, 1978) and that reside within and define environments (Endler, 1981). Researchers interested in the study of situations have focused on the symbolic, temporal, and frequently social nature of situations (Reis, 2008; Ten Berge & De Raad, 1999), although situations need not be restricted to purely social or interpersonal episodes (Sherman et al., 2010).

Similar to environments, situations can be distinguished in terms of objective and psychological characteristics. The objective, or nominal, situation is that as it exists outside the actor, and that can be defined in terms of physical or social variables (Ekehammar, 1974; Jessor, 1956; Saucier, Bel-Bahar, & Fernandez, 2007; Shoda, Mischel, & Wright, 2004). The psychological situation represents the situation as perceived by the actor(s), and that is defined in

terms of psychological variables (Edwards & Templeton, 2005; Magnusson, 1971). The psychological situation can further be distinguished on the basis of consensual features and individual features. Consensual, or canonical, psychological features correspond to shared representations (e.g., knowledge, concepts, beliefs, meanings; Block & Block, 1981; Saucier et al., 2007) associated with attributes such as standards, affective reactions, and appropriate or normative behavior for the self and others (norms, rules, and expectations; Cantor, 1981; Forgas, 1976; 1983; Schutte et al., 1985). Individual or subjective features pertain to aspects of the situation that are salient to the individual perceiver (Block & Block, 1981; Saucier et al., 2007). An example of the distinction between individual and consensual psychological features in the organizational psychology literature pertains to the delineation between psychological and organizational climate perceptions (e.g., James & Jones, 1974; Schneider, 1975).

The present study is restricted largely to psychological characteristics of situations. An understanding of objective characteristics of situations has applications in certain areas of SJT design (e.g., designing video-based or animated tests). However, psychological characteristics are argued to have wider applicability across simulation-based situational measurement procedures such as SJTs. Furthermore, in most instances, objective characteristics of situations are only meaningful insofar as they influence psychological characteristics of situations. Table 1 shows a representative sample of psychological features of situations from prior research in personality psychology[1]. The majority of the features shown in Table 1 were empirically derived

[1] For brevity, the individual studies shown in Table 1 are not reviewed here. As described in the Method section, the characteristics in Table 1 are applied in the development of an inventory of features, the Situational Characteristics Inventory, relevant for situations likely to be found in situational judgment tests used in organizational and educational contexts.

via factor-analytic, cluster-analytic, or other empirically-driven dimension-reducing methods, although there are exceptions (e.g., Block & Block, 1981; Reis, 2008).

Situational Characteristics and Behavioral Consistency

Given the diverse array of situational characteristics shown in Table 1, it is useful to consider whether such features are useful for explaining consistency in behavior across situations. The term "behavioral consistency" has been used in at least six ways in personality research (Fleeson & Noftle, 2008): (1) consistency in relative position /differential consistency: consistency in rank-order position for a single act; (2) aggregated correlational consistency: consistency in rank-order position for a composite or aggregate of behaviors; (3) coherence: consistency in the psychological underpinnings (e.g., cognition, affect, motivation) of behavior in spite of changes in observable or overt behavior; (4) ipsative consistency: consistency in an individual's configuration of behaviors irrespective of between-person, rank-order change (e.g., an individual being more agreeable than conscientious across two situations, even though her or his normative rank-order position on agreeableness and conscientiousness may increase or decrease over those situations; see also Furr & Funder, 2004); (5) temporal consistency: consistency in behavior over time across similar situations, and; (6) consistency of contingencies: consistency in terms of the manner in which situational contingencies affect changes in one's behavior (e.g., consistency in *if…then…* behavioral signatures; Mischel & Shoda, 1995; Shoda et al., 1994; Smith, Shoda, Cumming, & Smoll, 2009).

Early research on behavioral consistency used stimulus-response (S-R) inventories to examine the joint effects associated with persons, situations, and behaviors in accounting for response variance. S-R inventories sample situations relevant to a domain of interest (e.g., aggression, anxiety) and possible behavioral modes that individuals could enact in light of the

situations. In the typical S-R inventory, all facets (persons, situations, and behaviors) are crossed with one another (Endler & Hunt, 1966; 1968). Thus, all behaviors sampled are paired with all situations sampled and these combinations are presented to all respondents. Instructions prompt respondents to rate the intensity or the appropriateness of each behavior for each situation, with responses being interpreted as behavioral indicators of the trait in question (Endler & Hunt, 1968; Price & Bouffard, 1974).

A common finding of studies using S-R inventories is that variance attributable to the main effects of persons, situations, and behaviors is frequently less than or equal to that attributable to the simple interactions among these components (Ekehammar, 1974). In other words, the person-situation, person-response, and situation-response interactions together are as important in terms of predicting responses as are any of the main effects in isolation, as found in research on anxiety (e.g., Endler, 1966; Endler & Hunt, 1966), hostility (Endler & Hunt, 1968), and, more recently, the FFM personality traits (Van Heck, Perugini, Caprara, & Froger, 1994). In this context, the person-situation interaction component has been interpreted as evidence that respondents modify their behavior in light of the specific situations confronting them, whereas the magnitude of the situation-response interaction component indicates the extent to which the situations induce systematic variability in behavior across respondents (Endler & Hunt, 1966).

Collectively, research relying on S-R inventories and other similar methods provides evidence for systematic between-persons variability in the manifestation of trait-relevant behavior across situations. Although there is value in knowing the proportion of response variance attributable to each source (persons, situations, and behaviors), research using S-R inventories is silent with regard to specific variables associated with each source that account for response variance (Ekehammar, 1974). Therefore, researchers have attempted to isolate person

(e.g., self-reported consistency in behavior; Bem & Allen, 1974), situational (e.g., contingencies; Fleeson, 2007) and behavioral (e.g., perceived behavioral appropriateness; Price & Bouffard, 1974) characteristics that account for variability in responses. Of particular relevance for present purposes are situational characteristics and behavioral characteristics.

One situational characteristic that has been hypothesized as an influence on behavioral consistency is situational similarity. Behavioral consistency across situations should increase as situations become more similar or comparable to one another (Furr & Funder, 2004). Similarity may be defined on the basis of either subjective or objective terms. In two studies examining interpersonal situations, Furr and Funder (2004) found that rank-order consistency in behavior and consistency in behavior profiles were related to both subjective and objective situational similarity, with slightly more favorable results found for objective similarity. Similar findings were reported by Klirs and Revelle (1986) and Magnusson and Ekehammar (1978). Sherman and colleagues (2010) examined the relationship between situational similarity defined with respect to psychological features of situations, and behavioral consistency as assessed by the consistency of the behaviors reported by respondents. Ratings of the situational attributes were obtained from respondents. Situational similarity was related to behavioral consistency at both the between-person (r = .66 for the respondents' evaluations of situational features) and within-person levels (r = .63 based on the respondents' evaluations of situational characteristics).

Shoda and colleagues (1993) examined the extent to which functional similarity, defined on the basis of demands placed on the respondent (i.e., social, physical and motor, cognitive, self-regulatory), influenced consistency in verbally aggressive behavior among children in a New Hampshire summer camp. Shoda and colleagues (1993) found that situations varied considerably with regard to the types of demands imposed on the children and that consistency in aggressive

behavior was affected by the demands imposed by the situation. In another sample of children from the same New England summer camp, Shoda and colleagues (1994) examined patterns of cross-situational behavioral consistency as a function of two psychological characteristics of situations, namely the nature of the relationship with the other person in the situation (adult versus peer) and valence (positive versus negative). Similar to Furr and Funder (2004), they found that cross-situational consistency in various types of behavior (e.g., verbal aggression, compliance) was greater in situations that shared a larger number of common elements.

Finally, in two studies, Fleeson (2007) examined situation-based contingencies, or relationships between the manifestation of trait content in a given situation and psychological characteristics of that situation, for several FFM personality characteristics. In one study, the manifestation of behaviors associated with Extraversion, Agreeableness, and Conscientiousness were examined in situations that varied with regard to anonymity, friendliness, and task orientation. Extraversion was observed to a greater extent in situations high (as opposed to low) in friendliness, Agreeableness was observed to a greater extent in situations low in task orientation and high in friendliness, and Conscientiousness was observed to a greater extent in situations high in task orientation. In a second study, the manifestation of behaviors associated with Extraversion, Neuroticism, and Conscientiousness were examined in situations that varied with regard to anonymity, other's status, and task orientation. Extraversion was observed to a lesser extent in situations characterized by high task orientation and to a greater extent in situations characterized by greater status of others, Neuroticism was observed to a greater extent in situations characterized by high task orientation, and Conscientiousness was observed to a greater extent in situations characterized by high anonymity and task orientation.

The discussion thus far has focused on situational characteristics that influence consistency in behavior, with the intention of demonstrating that such characteristics are also likely to be relevant for describing the incidents present in SJT item stems, and that between-item variation in these characteristics may result in between-item variation in psychometric characteristics of SJTs. Research on behavioral consistency suggests, however, that variability in response depends not only on situational characteristics, but also characteristics of the behaviors themselves. Such research has implications for the individual response options used in SJTs, which is particularly relevant for tests that are scored at the response-option level (i.e., where responses are measured on a Likert-type scale for each individual response option).

Relative to the literature that has accrued on situational characteristics underlying behavioral consistency, there exists relatively little in the way of theory or empirical research on characteristics of behaviors associated with behavioral consistency. Price and Bouffard (1974) examined four properties of behavior that they argued influence the perceived appropriateness of potential responses: the extent to which the behavior elicits disapproval or embarrassment when performed outside its proper context, the extent to which other people would likely have second thoughts prior to engaging in the behavior, the extent to which someone else might report that the behavior in question is inappropriate irrespective of the situation or context, and the extent to which the respondent would say that the behavior is inappropriate irrespective of the situation or context. Other researchers have also examined the perceived appropriateness of behavior in various domains (e.g., social behaviors; Hill, 1989; Thompson, Royce, & Bankart, 1987).

Shoda and colleagues (1993) argued that some behaviors elicited in response to situational demands reflect relatively automatic or unmediated responses, whereas other behaviors are more cognitively mediated. Shoda and colleagues further suggested that cross-

situational consistency in cognitively-mediated behaviors should be less dependent upon the similarity of the situations with regard to demands imposed upon respondents than should consistency in unmediated or relatively automatic behavior. In support of this assertion, Shoda et al. (1993) found that cross-situational behavioral consistency in aggressive verbal behaviors (a relatively automatic response) was influenced by situational demands, whereas consistency in prosocial verbal behaviors (a cognitive-mediated response) was not.

Furr and Funder (2004) attempted to replicate the finding from Shoda and colleagues (1993). Furr and Funder instructed judges to rate a variety of behaviors that could be elicited in response to various situations on five-point Likert-type scales assessing the extent to which each behavior was cognitively mediated versus automatic. Using this approach, support was not found for a relationship between behavioral automaticity and the degree to which consistency depended on behavioral similarity. Thus, findings concerning the dependency of behavioral consistency on the automaticity of the behaviors in question are mixed. In addition to studying the automaticity of the behavioral responses, Furr and Funder obtained ratings on the individual behaviors with regard to their social desirability. Ratings of perceived automaticity and social desirability were weakly correlated (r = .09). Similar to what was found for automaticity, however, ratings of behavioral social desirability were not strongly related to the extent to which consistency in the behavior over situations was dependent upon situational similarity. In other words, the influence of situational similarity on cross-situational consistency in behavior did not appear to vary in a linear fashion as a function of the social desirability of the behaviors involved.

In summary, theory and research on behavioral consistency from the personality and social psychology literature suggests that psychological characteristics of both situations and behaviors influence the degree to which consistency in behavior is observed. If findings

concerning behavioral consistency generalize to SJTs and situation-based measurement procedures in general, one implication is that the psychological characteristics of the content used in SJTs likely influences psychometric characteristics of these tests. Prior to delving into this issue more directly, however, studies investigating situational characteristics in organizational research are first reviewed. This review highlights applications of situational concepts to problems of interest to applied research.

The Study of Situations in Organizational Research

Much of the theory and research on the psychological characteristics of situations originates in personality and social psychology. However, organizational researchers have also taken interest in the study of situations in workplace contexts. That there is a measurement method referred to as a *situational* judgment test arguably reflects organizational psychologists' awareness of the importance of situations in understanding workplace behavior. The research discussed in this section focuses on three areas of study related to situations in organizational research: (1) research on situational strength and constraints; (2) the concept of trait activation, and; (3) the study of behavioral consistency as it pertains to assessment centers (ACs).

The concept of situational strength is frequently traced to Mischel's (1973; see also Mischel, 1977) consideration of the conditions under which individual differences in personality characteristics should be most meaningful for predicting behavior. Mischel (1973) argued that situations affect human behavior to the extent that perceived situational characteristics impact cognition and affect that underlie behavior. Classes of relevant cognitive and affective constructs include encodings and personal constructs, expectancies and beliefs about the self and other objects, affects, goals and values, and skills, competencies, and self-regulatory strategies

(Mischel, 1973; Mischel & Shoda, 1995; 1998; Shoda, Mischel, & Wright, 1993). Situations are "strong" or "powerful" to the extent that they (a) invoke similar construals of the events that transpire across actors; (b) invoke uniform expectancies concerning behavioral appropriateness across actors; (c) provide adequate incentives to actors to behave as deemed appropriate, and; (d) instill the necessary skills for construction and execution of behavior (Mischel, 1973). Conversely, situations are "weak" to the extent that they induce variability in construals and expectancies among actors, afford insufficient incentives for performance, and place demands for responses for which at least some actors lack the necessary competencies or skills for adequate performance.

In weak situations, it is hypothesized that greater between-persons variance in behavior will be observed and that personality will exhibit its strongest influence on behavior; in strong situations, between-person behavioral variance is attenuated and the relevance of personality decreases (e.g., Cooper & Withey, 2009; Snyder & Ickes, 1985; although see also Marshall & Brown, 2006). There is some empirical evidence that situational strength acts as an influence on the degree to which personality predicts behavior (e.g., Marshall & Brown, 2006; Monson et al., 1982). However, in their recent review of the literature on situational strength, Cooper and Withey (2009) concluded that the overall body of empirical research supports neither the concept of strength, in its present conceptual state, or its hypothesized effects.

Given that situational strength is hypothesized to moderate relationships between individual difference characteristics and behavior (e.g., Mischel, 1973), researchers interested in personality in workplace settings have studied whether strength affects personality-outcome relationships. Recently, Meyer, Dalal, and Bonaccio (2009) conducted a meta-analysis of the moderating effect of strength on the conscientiousness-job performance relationship. Because

strength has been operationalized in numerous ways (e.g., situational ambiguity or uncertainty, task structure, industry norms, climate strength), Meyer and colleagues focused on two components of strength. The first component considered by Meyer et al. (2009) is constraints, defined as the extent of behavioral or decisional restriction placed on an employee through policies, procedures, government regulations and legislation, and so forth. The second component discussed by Meyer et al. (2009) is consequences, defined as the existence of contingencies between one's behaviors or decisions and outcomes that accrue to oneself, other employees, the organization as a whole, or external stakeholders.

Constraints, consequences, and overall strength were examined by Meyer and colleagues (2009) as moderators of the relationship between conscientiousness and both overall performance and task performance. Overall situational strength significantly moderated the conscientiousness-overall performance relationship, whereas the moderating effects of constraints and consequences were marginally significant. For task performance, constraints significantly moderated the relationship between conscientiousness and performance; the moderating effects associated with consequences and overall situational strength were both marginally significant. Thus, results provided some support for the moderating effects of strength in predicting overall performance, although the findings were not entirely conclusive.

Meyer, Dalal, and Hermida (2010) conducted a qualitative review of research on situational strength in the organizational sciences. Meyer et al. (2010) argued that strength can be represented by a four-facet structure: clarity (the degree to which cues concerning responsibilities and requirements are available and comprehensible), consistency (the degree to which cues associated with responsibilities and requirements are compatible with one another), constraints (the degree to which decision making and behavioral freedom are limited by forces

beyond the individual's control), and consequences (the degree to which actions or decisions have implications for other persons or entities). Meyer and colleagues (2010) discussed how each facet restricts the range of variability in behavior (e.g., clarity restricts the range of behavior by providing uniform information to all employees concerning behavioral expectations) and is influenced by different factors. One aspect that is not entirely clear in Meyer and colleagues' (2010) facet-based structure, however, is whether it should correspond directly to Mischel's original conceptualization of strength, as Mischel emphasized aspects (e.g., skills and competencies) that were not explicitly addressed by Meyer et al. (2010).

A related situational construct that has received attention in organizational research is situational constraints (Peters, Fisher, & O'Connor, 1982; Peters & O'Connor, 1980; Villanova & Roman, 1993). Peters and O'Connor's (1980) taxonomy of situational constraints included eight variables: job-related information, tools and equipment, materials and supplies, budgetary support, required services and help from others, task preparation, time availability, and work environment. The notion of situational constraints is similar to that of situational strength as originally defined by Mischel (1973; 1977) in terms of some of its predicted effects (e.g., restriction of behavioral variability) and in that it includes some consideration of cognitive-motivational variables (e.g., expectancies and beliefs). As mentioned, constraints have been discussed as a facet or operationalization of strength in recent reviews of the organizational literature (e.g., Meyer et al., 2009; 2010), although the two concepts seemed to have developed independently of one another. Peters and O'Connor (1980) hypothesized that situational constraints would have direct effects on outcomes (e.g., performance, motivation), would restrict variance in outcomes and correlations between other variables (e.g., ability, personality) and outcomes, would result in affective reactions such as frustration (particularly for highly

motivated employees), and that the removal of constraints should have both short- and long-term effects on performance.

Subsequent to Peters and O'Connor's (1980) publication, researchers examined the role of constraints with regard to the prediction of outcomes such as performance, affective reactions, and turnover (e.g., O'Connor, Peters, Pooyan, Weekley, Frank, and Erenkrantz, 1984), as well as the moderating effects of constraints on performance and affect-related outcomes (Peters, Chassie, Lindholm, O'Connor, & Kline, 1982; Peters, Fisher, & O'Connor, 1982). Villanova and Roman's (1993) meta-analytic review yielded estimates of -.14 and .21 for relationships between performance and turnover, respectively, and constraints, although each of the relationships was moderated by methodological features (e.g., constraint-performance relationships were stronger in lab than field settings; constraint-turnover relationships were stronger when turnover was operationalized with intent than actual turnover). Relationships with affect-related outcomes were somewhat stronger than those involving performance or turnover (job satisfaction, mean r = -.32; frustration, mean r = .39; commitment, mean r = -.22).

Other research examining situations in organizational settings has focused not on specific characteristics such as constraints or strength, but rather builds on research regarding situational characteristics and interactionism from personality psychology, social psychology and so forth in attempting to understand the manifestation of individual difference characteristics in behavior. For instance, Tett and colleagues (Tett & Burnett, 2003; Tett & Guterman, 2000) introduced the concept of trait activation, which implies that the manifestation of behavior associated with given trait content requires arousal through the presentation of situational cues.

Two concepts central to trait activation are situational strength, defined in accordance with Mischel's theoretical framework, and situation trait relevance, which describes the thematic

connection between the cues that define the situation and responses which indicate trait standing (Tett & Burnett, 2003). Trait relevance is argued to be a qualitative feature of situations that reflects which traits underlie behavior, whereas strength is a characteristic that resides along a continuum and that influences the degree of variability observed in trait-relevant behavior (Lievens, Chasteen, Day, & Christiansen, 2006). This distinction is comparable to that of Magnusson's (1981) categorization of structure characteristics, which reflect quantitative features of situations associated with complexity, clarity, strength, and promotion/restriction, and content characteristics, which include qualitative features of situations associated with specific tasks, rules, roles, goals, and so forth.

Building on the notion of trait activation, Tett and Burnett (2003) presented a model of job performance based on the propositions that traits are expressed in organizational settings in response to situational cues and that sources of cues exist at three levels: task, social, and organizational. Task cues reflect features that stem from the work performed within the position, including daily tasks, responsibilities and role requirements, and procedures. Social cues are features that arise from working with others in a social setting; these include needs and expectations of other parties, communication, behaviors that are socially prescribed, and team functions. Finally, organizational cues pertain to cues associated with organizational culture and climate.

Tett and Burnett (2003) distinguished between types of cues (demands, distracters, constraints, releasers, and facilitators) that cut across the task/social/organizational distinction. Demands are opportunities to act in a manner that is positively valued by the organization (e.g., formal and informal tasks and duties), whereas distracters are cues that promote opportunities to behave in negatively-valued ways (e.g., having access to the Internet that results in

procrastination during work hours). Constraints negate the influence of a trait on organizational behavior by restricting cues for trait expression (e.g., organizational policies banning the use of cell phones for texting during one's shift). Finally, releasers are cues that counteract constraints (e.g., one's supervisor taking frequent breaks so that she doesn't notice people texting during their shifts), whereas facilitators are cues that make the relevance of pre-existing trait information more salient. These varieties of cues can be further distinguished on the basis of activation status (demands, distracters, and releasers have a positive effect on trait relevance; constraints dampen the relevance of a trait; facilitators influence the activating or deactivating effects of other features), behavioral value (demands positively influence the value of trait-relevant behavior, whereas distracters negatively influence the value of trait-relevant behavior; constraints, releasers, and facilitators can have either a positive or negative effect), and frequency (demands, distracters, and constraints are chronic and ongoing, whereas releasers and facilitators are acute).

According to Tett and Burnett's (2003) model of performance, traits and situations directly influence workplace behavior and each of task, social, and organizational cues affect situation trait relevance (i.e., they moderate the effects of a given trait on performance). Trait activation theory has been applied to research on AC construct validity and the apparent paradox of ACs demonstrating consistent criterion-related and content validity evidence, but little in the way of construct validity evidence (Arthur, Day, & Woehr, 2008). In particular, although AC designers employ multiple exercises or tasks (e.g., in-baskets, leaderless group discussions, role plays) to elicit various behavioral dimensions from respondents, correlations of ratings of a given dimension across exercises (e.g., ratings of persuasion across a leaderless group discussion and a role play) are generally smaller in magnitude than are correlations of ratings of different

dimensions within an exercise (e.g., ratings of persuasion and empathy within a leaderless group discussion; e.g., Bycio, Alvarez, & Hahn, 1987; Highhouse & Harris, 1993; Sackett & Dreher, 1982; Sackett & Harris, 1988).

A number of arguments have been put forth as to why such a pattern of findings occurs with respect to AC construct validity. One argument relevant to the present discussion is that AC exercises, corresponding to different types of situations, vary with regard to their standing on psychological characteristics that influence the degree to which trait-relevant behavior is manifest (e.g., Neidig & Neidig, 1984; Sackett & Dreher, 1982; 1984). Consequently, researchers have examined the relationship between similarity across AC exercises in terms of psychological characteristics and consistency in ratings across exercises (e.g., Highhouse and Harris, 1993). Tett and Burnett (2003) suggested that consistency in dimension ratings across exercises should be predicted only under conditions where the exercises are associated with similar trait-relevant cues and when trait-relevant behaviors are valued equally across exercises (see also Lievens & Conway, 2001).

To this end, Haaland and Christiansen (2002) examined the relationship between consistency in ratings across exercises and the degree to which exercises afford the opportunity to observe trait-relevant behavior using the FFM personality traits. In order to assess the trait activation potential (TAP) of each exercise, raters judged the extent to which the exercises afforded the opportunity to observe a variety of behaviors associated with each FFM trait. Raters also judged each trait-exercise combination with respect to the extent to which the trait was relevant to the various exercises. Haaland and Christiansen (2002) found that ratings made on the basis of high-TAP exercises were more strongly correlated with self-report personality scores than were ratings made on the basis of low-TAP exercises. Furthermore, higher convergence in

behavioral ratings across exercises was found for high-TAP exercises (mean r = .30) than for low-TAP exercises (mean r = .15).

Similar to Haaland and Christiansen (2002), Lievens and colleagues (2006) suggested that convergence in same-dimension ratings across exercises should be poorer when the exercises differ with regard to their activation potential for the trait or dimension in question. Conversely, stronger convergence in ratings is expected when the ratings are derived from exercises where there is a greater degree of opportunity to observe trait-relevant behavior. Furthermore, because seemingly distinct behaviors or dimensions may represent expressions of a common underlying trait (e.g., communication and dominance both reflecting extraversion), discrimination among dimension ratings within exercises should be poorer for behaviors that are manifestations of a single trait. Relevant to the present discussion, Lievens et al. (2006) found some support for the arguments that (a) ratings from high-TAP exercises would exhibit greater convergence than ratings from low-TAP exercises, and (b) greater discrimination among ratings within an exercise would be exhibited for behaviors not representing manifestations of a common underlying trait.

THE PRESENT STUDY

Researchers have called for a greater understanding of the characteristics of SJT content (e.g., Weekley and Jones, 1999). For instance, McDaniel and Nguyen (2001) observed that characteristics of SJT items vary widely across tests; consequently, they called for additional research on item characteristics influencing validity. To this end, the proposed study applies concepts from research on interactionism and the psychological features of situations to the study of situational characteristics in SJTs with the ultimate intent of contributing to the field's understanding of the psychometric characteristics of SJT scores (i.e., correlations with external variables and criterion-related validity).

Some research in this vein has recently been conducted, primarily with regard to ACs (e.g., Haaland & Christiansen, 2002; Highhouse & Harris, 1993; Lievens et al., 2006). There have also been initial attempts to systematically examine the content of SJTs. For instance, Kell et al. (2010) examined the content of critical incidents used in SJTs by having research assistants rate the extent to which the incidents described an interpersonal situation and the extent to which the incidents described a task situation. Kell and colleagues' (2010) distinction between task and interpersonal cues is congruent with the broad categories of task- and social-level cues in Tett and Burnett's (2003) discussion of situational cues affecting trait relevance and Johns' (2006) dimensions of discrete context. However, there are two related limitations of rating or categorizing item stems or incidents using the task/problem-solving versus social/interpersonal/helping distinction: (1) it ignores many situational features or cues that have been delineated in the personality, social, and organizational literatures, and (2) it may be too gross or too broad for effectively differentiating SJT items in terms of psychometric characteristics. Tett and Burnett (2003) illustrated how various levels and varieties of situational

cues beyond the task-social distinction may influence trait relevance, which suggests that analyses of situational content at a finer level of detail than Kell and colleagues' distinction between task and interpersonal may be merited. The present study marks an initial attempt at sampling and examining an extensive array of situational features that comprise SJT item content.

The proposed study entails two primary components. The first component is the collection of item-level information regarding SJT content as it pertains to the situational features inherent in SJT item stems and the behavioral features inherent in SJT response options. Once these data are collected, the second component entails the linkage of SJT situational and behavioral characteristics to psychometric characteristics of SJT scores in terms of relationships with other individual difference variables and criterion-related validity.

With respect to situational and behavioral characteristics of SJT content, the focus will be on psychological features as opposed to nominal or objective features, given that individuals perceive situations primarily in terms of psychological as opposed to nominal attributes (e.g., Forgas, 1976; 1983). Additionally, knowledge of the relevance of psychological characteristics for psychometric properties of SJTs is arguably more generalizable than information derived from an analogous investigation of nominal or physical characteristics. Asking someone on a date for the first time, being on one's first job interview following college graduation, and performing a saxophone solo in a high school concert represent nominally diverse situations. Psychologically, however, these situations share certain features (e.g., anxiety provoking, felt sense of evaluation, heightened concern with rejection) that may be relevant for understanding how responses to these situations are associated with various personality characteristics or how such responses can be used to predict theoretically-relevant criteria. This commonality would be

captured by relevant psychological characteristics associated with performance anxiety or evaluation, but would be hidden by a focus on nominal or surface-level characteristics.

For several reasons, interest is also in consensual or canonical features as opposed to individual-specific or idiosyncratic features. First, one motivation underlying the proposed study is an understanding of design characteristics as properties of SJT content as opposed to idiosyncratic perceptions that reflect properties of individuals (e.g., respondents). In an operational setting, individual-specific or idiosyncratic features cannot be known by developers until the respondent is being administered the test. Similarly, a test developer cannot, and arguably will never, know all respondents' idiosyncratic perceptions during the process of reviewing item content. Finally, calls have been made to advance the field's understanding of abstract psychological features of situations independent of individuals' construals (e.g., Reis, 2008; Wagerman & Funder, 2008). To be clear, none of these arguments implies that knowledge of idiosyncratic features is irrelevant or uninteresting; indeed, the meaning that individuals assign to situations is, in some respects, both shared and unique (Fournier, Moskowitz, & Zuroff, 2008). Along these lines, a thorough understanding of the influence of item design characteristics requires knowledge in two areas: (1) the influence of test stimulus properties on psychometric outcomes, such as item difficulty, and; (2) the process or processes used by respondents in responding to the task presented during the assessment process (Enright, Morley, & Sheehan, 2002). Interest in the present study is primarily in the former; that is, in understanding characteristics of situations and behaviors as features of SJT item content and how such features relate to item- and test-level outcomes (see also Funder, 2008).

Thus, to recapitulate, the primary motivation for the proposed study is that a finer-grained analysis of the properties of item- and test-level SJT data can be undertaken by incorporating

knowledge regarding the psychologically active characteristics of SJT item content. This argument is in accord with shifts in the conceptualization of personality from de-contextualized, global characteristics to patterns of behavior conditional upon situational characteristics (Mischel & Shoda, 1998). This argument is also in agreement with recent conceptualizations of evidence-based test design principles that emphasize the importance of understanding properties of the task stimuli used in complex assessment procedures (e.g., Mislevy, Almond, & Lukas, 2003; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002).

When scoring SJTs designed to assess a given characteristic, test developers generally propose aggregating response scores across item stems or situations. Such a scoring protocol treats situations as parallel indicators of the characteristic of interest, similar to arguments made with regard to AC exercises (Haaland & Christiansen, 2002). Consequently, variability within respondents across situations in situation-based measurement procedures (item stems for SJTs, exercises for ACs) is viewed as error. Treating within-person variability as error motivates the use of aggregation to circumvent error or specificity in order to obtain a better approximation of the construct or latent trait in question (e.g., Epstein, 1980; 1983; Epstein & O'Brien, 1985; Mischel, 2004; Mischel & Shoda, 1995). Some amount of within-person variation in behavior in situation-based measurement procedures will reflect error as commonly defined. However, it seems rather implausible that all within-person response variation in situation-based measurement procedures is error. Rather, if personality research on situations, interactionism, and behavioral consistency generalizes to SJTs and ACs, then some portion of within-person variability in responding reflects substantively meaningful variation that is lost when responses across situations are aggregated. If true, situations sampled in situation-based measurement

procedures cannot simply be assumed to represent parallel indicators of a given construct of interest (Bledow & Frese, 2009; Haaland & Christiansen, 2002).

A first step in understanding behavioral consistency, or the lack thereof, across situations is an understanding of the psychologically relevant features of situations (see also Mischel, 2004; Mischel & Shoda, 1995; 1998). In this respect, the proposed study is consistent with the argument that the assessment of individuals in context necessitates understanding of the psychologically active features of situational content that define the stimuli being used (e.g., Cervone, Shadel, & Jencius, 2001). Indeed, inconsistency in behavior may become predictable when one arrives at an understanding of the psychological characteristics of situations in question. Understanding psychological features of situations may permit the identification of what Fleeson and Noftle (2008) refer to as "regions of local consistency." As Fleeson and Noftle note, it is not reasonable to assume that uniform consistency will be present over situational and behavioral content; rather, there are likely to be regions of situations or behaviors that exhibit greater consistency. If psychological features of situations shed light on groups of psychologically equivalent situations, this would arguably represent an advance in SJT design.

Similarly, Bledow and Frese (2009) stated that one of the strengths of SJTs is their ability to measure persons in situ. However, Bledow and Frese also note that respondents take the particulars underlying the situation into account when deciding upon an appropriate response. Implicitly, SJT designers likely acknowledge that respondents focus on specific characteristics of item stem content in making response choices; generally, then, one seeks to sample broadly from the domain of feasible situations and behaviors, conditional upon a set of fixed environmental or contextual features (e.g., designing an SJT for use in, say, a manufacturing versus office environment). However, because respondents are unlikely to react in a uniform manner to all

situational stimuli in an SJT, test developers would benefit by better understanding how and why situations in SJTs differ from each other and how those differences influence measurement characteristics of SJTs. The hypotheses of interest for the present study are described below. In order to provide a concrete illustration of how the study would be conducted, Table 2 shows a subset of a hypothetical dataset that is referred to when discussing the study hypotheses. The reader will be referred to Table 2 where appropriate.

Study Hypotheses

Agreement and Reliability of Ratings of Psychological Characteristics of Situations and Response Options

In the present study, SJT content will be scaled according to its standing on various characteristics of interest. Scaling of content will be accomplished by having raters judge the applicability of each characteristic for each item stem or response option. Two prerequisites for examining rated psychological characteristics of the content in SJTs are consistency across raters or judges in ratings (i.e., inter-rater reliability and agreement; LeBreton & Senter, 2008) and variability across item stems or response options in ratings (i.e., the existence of between-stem or between-option differences in characteristics). In other words, judges' ratings of characteristics should be in agreement, and there should exist sufficient variability between item stems and response options in ratings to justify the use of ratings as predictors in models used to explain psychometric properties of SJT response data.

Consistency across raters has been found in examinations of the content in both ACs and SJTs. For instance, Haaland and Christiansen (2002) reported an average correlation of .72 between four raters who judged the extent to which AC exercises afforded the opportunity to

observe trait-relevant behaviors and the extent to which various behaviors had the potential to be observed across exercises. Using the same scales and six raters, Lievens et al. (2006) reported an estimate of .58 for Kendall's coefficient of concordance and an ICC of .90. Motowidlo and Beier (2010) reported an average correlation of .53 between six raters with regard to the extent to which SJT response options indicated expressions of Agreeableness, Conscientiousness, and Extraversion. For two groups of seven raters, Kell et al. (2010) reported inter-rater reliability estimates in the upper .80s and .90s for ratings of personality trait expression associated with the behaviors taken in response to various critical incidents.

With respect to the existence of between-stem or between-option differences in rated characteristics, researchers examining ACs and SJTs have reported that exercises or situations vary in terms of the relevant characteristics on which they are rated. For instance, Tett and Guterman (2000) obtained trait-relevance ratings for five different types of situations (e.g., risk taking, complexity, sociability) on four-point scales. Ten situations for each of the five types were examined (e.g., ten risk-taking situations, ten complex situations) for a total of 50 situations. For each of the five situation types, the ranges in the mean scores across the ten situations were 2.72, 2.37, 0.93, 2.25, and 1.53, which are non-negligible given that the maximum possible range was four. Across 100 critical incidents, Kell et al. (2010) prompted SMEs to rate the FFM trait expression associated with behaviors taken in response to critical incidents for two jobs on seven-point scales. For each of the FFM traits, *SD*s for mean ratings ranged from 1.5 to 2.2 for the first job (100 critical incidents) and from 0.9 to 1.6 for the second job (97 critical incidents).

The studies described in the preceding two paragraphs demonstrate two points. First, raters' assessments of the content used in situational measures such as ACs and SJTs are capable

of attaining adequate reliability or consistency. Second, rated characteristics vary across the relevant unit under study, which for the present case suggests that between-stem and between-option variance in ratings will be found. As an example, if raters were asked to rate the degree to which a series of stems reflected time urgency, one would expect that (a) raters would be relatively consistent in their ratings of time urgency, such that ratings of time urgency for a given situation should be relatively similar in magnitude across raters, and (b) ratings of time urgency vary between stems, or that different item stems reflect varying levels of time urgency.

The implication of inter-rater reliability and agreement for Table 2 is that, for a given situational characteristic item, variance across stems will be large enough relative to variance across raters to justify aggregation. The implication of variance between stems or response options in characteristics is largely just that – when ratings are pooled across judges, situations will be differentiated on the characteristics of interest, thus permitting the analysis of between-stem or between-response option differences as predictors of SJT psychometric characteristics. Thus, in Table 2, values for the variables corresponding to stem characteristics (*stem_f1* through *stem_f5*) vary across situations. For *stem_f1*, stem 1 had a mean rating of 3.15 averaged across raters, whereas stem 8 had a mean rating of 6.52 across raters. These points pertain not to hypotheses that are theoretically or substantively central for the present study and thus will be posed as research questions that will be addressed as a prerequisite for conducting inferential procedures associated with the substantive issues of interest. Thus:

> *Research Question 1. Do ratings of psychological characteristics of SJT content demonstrate meaningful levels of inter-rater reliability and agreement?*
>
> *Research Question 2. Do ratings of psychological characteristics of SJT content demonstrate between-item variance?*

Psychological Characteristics of Situations and Correlations with Other Individual Difference Characteristics

Researchers have lamented the field's lack of understanding concerning SJT construct validity (e.g., Christian et al., 2010). One reason for studying the psychological features underlying SJT content is to approach the investigation of SJT construct validity from a theoretically-based perspective, in this case, a perspective rooted in behavioral consistency and interactionism. If applicable, psychological features inherent in SJT stems and response options in SJT content may be useful for understanding findings that have emerged with respect to convergent validities between SJTs and theoretically-relevant variables, namely those associated with constructs in the domains of personality, cognitive ability, experience, or knowledge. In addition to theoretical understanding, psychological features inherent in SJT stems and response options in SJT content may be applicable for predicting psychometric characteristics, which may be relevant for various practical applications (e.g., item development, computer-adaptive test administration, development of item pools for parallel test form development).

The arguments in the preceding paragraph are supported by theory and research on trait activation (e.g., Haaland & Christiansen, 2002; Lievens et al., 2006; Tett & Burnett, 2003; Tett & Guterman, 2000) that suggests that psychologically similar situations, where similarity is defined on the basis of trait relevance, are likely to be associated with behaviors that more similarly correlate with the relevant individual difference characteristic in question. In other words, correlations between item-level stem scores and individual difference variables should vary systematically as a function of the psychological features of the situation. For instance, time urgency, as a psychological feature associated with item stems, may differentiate stems in terms of their correlations with Neuroticism or Conscientiousness; as the time urgency of the situation

increases, correlations between item stem scores and trait Neuroticism or Conscientiousness may also increase.

Table 2 illustrates a hypothetical data structure for investigating relationships between stem characteristics and convergent validities for stem scores. Again, values associated with *stem_f1* through *stem_f5* reflect scores for each item stem on psychological features. For the present hypothesis, the other relevant variables in Table 2 are *r_cons* through *r_gma*, which pertain to zero-order correlations between item stem scores and individual difference characteristics. For present purposes, these correlations represented by *r_cons* through *r_gma* can be taken as indicators of what might be called *trait saturation* (e.g., saturation with cognitive ability, saturation with extraversion, etc.). Investigation of the relationship between psychological characteristics of item stems and trait saturation entails the examination of zero-order correlations or regressions between stem standing on various characteristics of interest (i.e., *stem_f1* to *stem_f5* in Table 2) and correlations between stem scores and individual difference characteristics (i.e., *r_cons* to *r_gma* in Table 2). Theoretically, then, psychological characteristics of SJT stem content will differentiate SJT responses in terms of indices of association (e.g., zero-order correlation coefficients, regression coefficients) between stem scores and other person characteristics. Specifically:

> *Hypothesis 1. Between-stem variability in trait saturation (i.e., correlations between stem scores and personality characteristics, cognitive ability, experience, knowledge) will be accounted for by situational characteristics associated with stems. Thus, non-zero correlations will be observed between (a) situational characteristics of item stems and (b) stem trait saturation.*

Psychological Characteristics of Situations and Correlations with Relevant Outcome Variables

If psychological features of situations are systematically related to item stem trait saturation as argued above, it logically follows that psychological features of situations should

affect the extent to which stem scores correlate with theoretically-relevant outcome variables such as performance, attitudinal outcomes (e.g., satisfaction), or measures of withdrawal. Specifically, if psychological features of situations influence the degree to which responses are indicative of specific individual difference characteristics (e.g., as per trait activation theory), then situations permitting expression of traits relevant to specific outcomes should yield stem scores that are more strongly related to the outcomes that those traits are associated with. For example, if one is using an SJT containing item stems that vary in the provision of cues that demand detail orientation (e.g., perhaps situations where the importance of providing a high-quality product is emphasized), and if the criterion being predicted is sensitive to variability across individuals in detail orientation, then situations that provide stronger cues relevant to detail orientation may yield responses that are more strongly associated with the criterion of interest.

With respect to Table 2, entries associated with the variables *r_perf*, *r_sat*, and *r_absent* reflect zero-order correlations between scores for each item stem and performance, satisfaction, and absenteeism. Investigation of the relationship between psychological characteristics of item stems and item stem criterion-related validities entails the examination of zero-order correlations and regressions between stem standing on various characteristics of interest (i.e., *stem_f1* to *stem_f5* in Table 2) and correlations between stem scores and outcome variables (i.e., *r_perf*, *r_sat*, and *r_absent* in Table 2). If psychological features influence the validity of item stems, then significant correlations or regressions would be expected between psychological features of the item stems (*stem_f1* through *stem_f5*) and correlations between the item stem scores and relevant outcomes (i.e., *r_perf*, *r_sat*, and *r_absent*).

Hypothesis 2. Between-stem variability in relationships between SJT stem scores and outcome variables will be accounted for by situational characteristics associated with

stems. Thus, non-zero correlations will be observed between (a) situational characteristics of item stems and (b) item stem criterion-related validities.

Joint Consideration of Situational Features of Item Stems and Behavioral Features of Response Options: An Interactionist Perspective

Up to this point, the study hypotheses have pertained to the psychological features of stems as situations and their associated effects on SJT psychometric outcomes. One justification for focusing on this level of analysis is that many SJT scoring protocols yield stem-level scores; that is, for a given item stem and associated series of response options, scores are frequently summed across response options within each stem to compute a stem-score (e.g., Ployhart & Ehrhart, 2003; Table 1, p. 4), as discussed earlier. Another justification for focusing on stems is the large body of personality and social psychological research cited above that demonstrates the relevance of situational features for understanding behavioral consistency.

However, the situation or stem is not the only relevant level of analysis when considering test or stimuli content used in situation-based measurement techniques. For instance, SJT scoring protocols frequently emphasize individual response option scoring. As an instance of a scoring format frequently used with "Should Do" instructions, Ployhart and Ehrhart (2003) noted that respondents are often instructed to rate the effectiveness of each of the individual response options in light of the common item stem. In this case, each response option is rated on, for instance, a five-point scale in terms of how effective the respondent evaluates each behavior in light of the situation.

This type of format shifts the focus from understanding stem score-level psychometric properties to understanding response option-level psychometric properties. To illustrate this type of scoring protocol, Table 3 provides an example item drawn from the Calibrator scale in Mumford, van Iddekinge, Morgeson, & Campion's (2008) Team Role Test. In addition to

scoring techniques, standardization inherent in SJTs as a measurement technique is incorporated partly through the presentation of specific behavioral response options to respondents (e.g., as opposed to an open-ended response format). When respondents are instructed to rate each individual response option, as in the example Team Role Test item in Table 3, each of the individual behaviors likely has features associated with it that, in light of the situation, will affect the psychometric features of the response option.

These arguments imply that, in certain circumstances, it is necessary to consider not only situational features associated with item stems, but also behavioral properties associated with the individual response options. Recognition of behavioral characteristics in light of situational features is certainly not novel to the present study. Research on behavioral consistency in the 1960s and 1970s using S-R inventories revealed that behavior main effects, as well as behavior-situation and behavior-person interactions, account for non-negligible response variance (e.g., Endler & Hunt, 1966; 1968; 1969). More recently, researchers interested in interactionism in personality and organizational psychology have examined specific features of behaviors, including how the behaviors in question can be scaled along dimensions such as automaticity (e.g., Furr & Funder, 2004; Shoda et al., 1993) and trait expression, or the degree to which the content associated with a given behavior is indicative of some specific underlying characteristic or trait (e.g., Motowidlo & Beier, 2010). For the present study, interest in behavioral characteristics will also pertain to the trait expression of the response options, given that (a) research on the automaticity of behaviors has not been consistent (Shoda et al., 1993 found evidence for automaticity, whereas Furr & Funder, 2004 did not) and (b) FFM trait expression provides a relatively comprehensive framework from which to consider the behavioral content of SJT items.

Whereas behavioral characteristics of response options (e.g., trait expression) may be useful for explaining psychometric outcomes of response option scores, the influence of behavioral characteristics may also depend on the situational characteristics of the item stem in question. In other words, specific behavioral features of response options (e.g., the extent to which the behavior in question is seen as reflecting dominance or surgency versus warmth or empathy) will have a greater influence on the psychometric properties of response option scores as a function of the situational features associated with the item stems. This argument is in accord with propositions set forth in trait activation theory, which suggest that cues inherent in situations influence the relevance of specific traits in the situation under question.

Another way of thinking about how situational (stem) features might influence the effect of behavioral (response option) features on psychometric outcomes comes from viewing SJTs as being composed of item bundles (Rosenbaum, 1988) or testlets (e.g.,Wainer & Kiely, 1987; Wang, Bradlow, & Wainer, 2002), both of which are frequently used to refer to clusters of items that are administered around a common stimulus (DeMars, 2006; Wainer & Thissen, 1996). Testlets are frequently discussed in the context of reading comprehension tests designed with groups or subsets of items that refer to common reading passages (Tuerlinckx & De Boeck, 2004). In the context of SJTs, testlets are formed around item stems, suggesting that the various behavioral response options are clustered around the stems, irrespective of the scoring format (i.e., a forced-choice protocol where examinees are prompted to choose the best and the worst response option, the most and least likely response option, etc.; a rating-scale protocol where examinees are prompted to rate each response option). Thinking about SJTs from a testlet perspective makes sense from both a design standpoint (viewing each item stem and series of response options as a bundle of items that are presented together) and from the psychological

perspective that, as mentioned above, respondents consider details of the situation described in the stem when choosing a response (Bledow & Frese, 2009). These considerations bolster the argument that response characteristics should be considered as conditional upon stem characteristics.

In summary, although the primary focus of the present research is to examine situational characteristics associated with the psychometric properties of SJTs, the use of certain scoring protocols as well as the design of SJTs around standardized behavioral response options necessitates consideration of the behavioral features of the response options. Therefore, trait expression underlying the behaviors represented in response options is a key concept in understanding SJT item content, in line with research on SJT design (e.g., Kell et al., 2010; Motowidlo & Beier, 2010), as well as measurement research concerning behavioral indicators of FFM characteristics (e.g., Jackson, Wood, Bogg, Walton, Harms, & Roberts, 2010). Thus, Hypotheses 1 and 2 are extended by suggesting that situational characteristics of stems and behavioral characteristics of response options interact in the prediction of psychometric outcomes of SJT response option-level scores (i.e., consistency in responses, correlations with person characteristics, and correlations with relevant criterion variables). More specifically:

> *Hypothesis 3a. Between-option variability in trait saturation will be accounted for by the interaction between situational characteristics of the stems and behavioral characteristics of the response options. Thus, situational characteristics of item stems and behavioral characteristics of the response options interact in the prediction of response option trait saturation.*
>
> *Hypothesis 3b. Between-option variability in criterion-related validity will be accounted for by the interaction between situational characteristics of the stems and behavioral characteristics of the response options. Thus, situational characteristics of item stems and behavioral characteristics of the response options interact in the prediction of response option criterion-related validity.*

To illustrate how Hypotheses 3a and 3b will be investigated, Table 4 shows a hypothetical dataset of situational characteristics of item stems, behavioral characteristics of response options in terms of trait expression, and estimated zero-order correlations between response option-level scores and person characteristics. To conserve space, only three situational characteristics are shown in Table 4 (*stem_f1* through *stem_f3*) and only Agreeableness, Conscientiousness, and Extraversion are shown in Table 4 (for both trait expression and correlations with person characteristics); in the actual analysis, however, all relevant variables would be found in the dataset. The data are presented in a stacked format. In other words, entries are repeated for each *stem* for each response option involved in each stem cluster. In this hypothetical illustration, each stem in the test is associated with five response options: response options a, b, c, d, and e. Although five response options cluster around each stem in this example, the number of response options is permitted to vary across stems in theory. Therefore, there are five entries for each stem; one row per response option.

Because entries for *stem* are repeated for each of the response options that cluster around the stem, entries for situational characteristics *stem_f1* through *stem_f3* also repeat across all of the entries for each stem. However, because there are different response options that cluster around each stem, entries for the behavioral characteristics associated with the response options (*ro_agr*, *ro_con*, and *ro_ext* in Table 4) vary across the response options within each stem. Thus, for stem one, response option (a) had trait expression scores of 1.96, 6.73, and 4.65 on Agreeableness, Conscientiousness, and Extraversion, respectively, whereas response option (b) had trait expression scores of 3.09, 5.90, and 1.69 on Agreeableness, Conscientiousness, and Extraversion, respectively. In order to test Hypotheses 3a and 3b, each of the relevant correlation coefficients would be modeled with an equation represented by an intercept term, the main

effects associated with the specific situational characteristic (stem) and trait expression (response option) variables in question, the interaction term between the situational and behavioral characteristic, and an error term. Significant effects associated with behavioral characteristics of response options will provide evidence that the trait expression associated with response option scores systematically influences correlations with either other individual difference characteristics or criterion measures. Significant effects associated with behavioral-situational interactions will indicate that the effects of trait expression will be contingent upon specific psychological features of the item stems in question.

METHOD

In order to address the hypotheses discussed in the last section, two sets of data relevant to common SJT material must be available. The first dataset will contain variables required to address Hypotheses 1 and 2, namely, data concerning situational characteristics corresponding to the item stems as provided by raters or judges, data on correlations between stem scores and external individual difference characteristics (from which to examine convergent and discriminant validity; Hypothesis 1), and data on correlations between stem scores and criterion outcomes relevant to the SJT in question (from which to examine criterion-related validity; Hypothesis 2). The second dataset will contain variables required to address Hypotheses 3a and 3b, namely, data concerning both situational characteristics corresponding to item stems as well as behavioral characteristics corresponding to the response options, data on correlations between response option scores and external individual difference characteristics (from which to examine convergent and discriminant validity; Hypothesis 3a), and data on correlations between response option scores and criterion outcomes relevant to the SJT in question (from which to examine criterion-related validity; Hypothesis 3b). The ratings of item stems and response options will be collected specifically for this study; the data pertaining to correlations with external characteristics and criterion measures will be obtained from archival sources. Each of these datasets structures will be described in greater detail below.

Data and Procedure

Perceived Situational Characteristics

Sampling raters. Similar to previous research examining psychological features of situations in both the personality and organizational literature, ratings were collected from

undergraduate students (as was done by Bergman et al., 2006; Furr & Funder, 2004; Kell et al., 2010; Lievens et al., 2006; Magnusson, 1971; Motowidlo & Beier, 2010; Study 1 and 2 in Motowidlo et al., 2006; Tett & Guterman, 2000). Using past research to guide the number of required raters is made difficult by the large variability in number of raters used across studies. Among the studies just cited, number of raters ranged from two (Kell et al., 2010; Motowidlo et al., 1990) to 438 (Motowidlo & Beier, 2010). Two studies employed 100 or more raters; namely, Tett & Guterman (2000) used 123 raters for judgments of trait relevance, whereas Motowidlo & Beier (2010) used 438 raters for judgments of the effectiveness of various behaviors). Aside from these two studies, the other studies cited above employed between two and 34 raters.

Two practical constraints limited the number of raters that could be utilized to provide ratings for the present study. First, the number of situations to be rated was large. Assume a fully-crossed design where raters evaluate all situations on all situational characteristics, and where ratings of psychological features are made on a ten-item instrument. With, say, 100 situations to rate, such a design would require that each rater provides 1,000 ratings, which is unrealistic. Second, the number of characteristics on which situations and behaviors will be rated is also non-negligible. As described below, a 43-item inventory was developed to assess the situational characteristics of interest. With 100 situations and 43 items, each rater would have to make 4,300 ratings for the stem characteristics alone, again a burden that is unrealistic.

In order to mitigate the obstacles associated with rater burden, a nested design was used where raters were responsible for rating only a subset of the total number of situations. Table 5 provides an illustration of the study design with 25 situations and 50 raters. Assuming again that there are 100 situations to rate in total, situations were divided into blocks of five, resulting in 20 total blocks (first block: situations 1-5, second block: situations 6-10, …, twentieth block:

situations 96-100). Each rater was assigned one block of situations; thus, raters were prompted to rate five situations. The same blocking approach was taken to obtaining ratings for the behavioral characteristics of the response options. Given the variability observed in the studies cited above with regard to number of raters and given that the situational characteristic inventory used in the present study has not previously been applied in empirical research, a pilot study was first undertaken to estimate the number of raters required to achieve adequate inter-rater reliability. The pilot study is elaborated upon below following the descriptions of the situational characteristic and behavioral characteristic inventories.

The situational characteristic inventory (SCI). Two approaches can be taken in studying the relevance of situational features for understanding the psychometric characteristics of SJT scores (Shoda, 2003). One approach, a *deductive* strategy, begins with specific constructs of interest derived from prior theory, intuition, or informal observation. The second approach, an *inductive* strategy, is largely exploratory in nature, seeking to discover psychological features that differentiate situations in terms of psychometric characteristics. Although there are specific constructs that have been examined in terms of situational or contextual features in the personality and organizational literatures (e.g., situational constraints, situational strength), reviews of these constructs have not yielded an unequivocal portrayal of their relevance (e.g., Cooper & Withey, 2009). Similarly, researchers in both the personality and organizational literatures have lamented the lack of a coherent, agreed-upon taxonomy of situational characteristics.

Given these considerations, the approach espoused within this study was somewhat of a *hybrid deductive-inductive* strategy toward sampling potential relevant situational features. The approach was *deductive* in that empirical research on situational characteristics was leveraged to

generate a sample of specific situational characteristics that served as the focus of examination. Because there is no theoretical justification for favoring specific features as there is no preferred theoretical taxonomy, because discrete situations are likely to be multidimensional in terms of their relevant features (e.g., Zayas & Shoda, 2009), and because the intent of the present study was to study and demonstrate the relevance associated with psychological characteristics of situations broadly as opposed to examining specific situational characteristics, situational characteristics will be sampled broadly and inclusively. The approach was inductive in that an attempt was made to delineate characteristics in terms of their relevance for understanding psychometric properties of SJT response data.

Inventories have been developed to assess psychological features of situations (e.g., the Riverside Situational Q-Sort; see Sherman et al., 2010). However, many of these inventories were not developed to sample the specific types of psychological features likely to be relevant to either organizational or educational contexts. Thus, a measure referred to as the Situational Characteristic Inventory (SCI) was developed for the present study based on the following approach. First, a thorough review of the literature on the psychological features of situations was conducted in the areas of personality and social psychology. As mentioned above, this literature pertained to the related topic areas of interactionism, behavioral consistency, and situational characteristics. The dimensions found in these studies, and the measures used to assess situations from these studies, were collected. A representative list of studies derived from this review is presented in Table 1, along with the dimensions derived from these studies.

Second, information obtained from the personality and social psychological literatures was complemented with constructs relevant to the description of situations that are specific to applied settings, particularly organizational contexts. Situational features relevant to applied

settings were drawn from two areas: work analysis and job design. Regarding the former, descriptors were drawn from the O*NET Work Context content model for 27 characteristics (e.g., consequences of error, impact of decisions on co-workers or company results, coordinate or lead others, responsibility for outcomes and results). Regarding the latter, Morgeson and Humphrey's (2006) recent review of the work design literature was consulted, with dimensions drawn from their model of motivational (e.g., autonomy, task identity), knowledge (e.g., job complexity, information processing), and social (e.g., social support, interaction outside the organization) work design features.

Third, the dimensions and item content collected above were compiled and reviewed thoroughly to delineate patterns, areas of agreement or congruence between the personality/social and organizational perspectives, and areas where constructs from the organizational psychology literature could be used to complement what was found in the personality and social psychology literatures. In terms of specific measures examined, these included the following: Battisch and Thompson's (1980) behavioral scales, affect scales, and situational descriptors; Eckes' (1995) 18-item inventory of situational descriptors; Fleeson's (2007) situational descriptor scales from studies 1 and 2; Forgas' (1976) 12-item inventory of situational descriptors; Haaland and Christiansen's (2005) 25-item scale; Morgeson and Humphrey's (2006) 63-item work design questionnaire; O*NET Work Context 56-item scale; Price and Bouffard's (1974) four-item scale assessing situational constraints, and; the Riverside Situational Q-Sort (RSQ) v. 3.15, an 89-item inventory of situational descriptors. In addition to including item content from these measures, situational descriptors were culled from Tett and Burnett's (2003) Table 2, which shows exemplar situational descriptors associated with task,

and, and organizational-level cues connected with the FFM, and Mumford and colleagues' (2006) discussion of situational cues relevant to team contexts.

Fourth, multiple theoretical domains of item characteristics were retained (Table 6). Domains that were included were those that were not redundant with other domains and that would be relevant to SJT content. In some instances, minor changes were made to the domain or definition to make it relevant for present purposes (e.g., the O*NET Work Context dimension include "Frequency of Conflict Situations," which was changed to "Conflict Situations"). Slight changes were also made to definitions in certain cases where a given characteristic (e.g., an O*NET Work Context feature or a work design feature) did not have an isomorphic analog relevant to the description of situations. The dimensions in the Customized/Adapted category were included because they appeared relevant to describing content of situations in SJTs, but were not represented in other domain categories listed above. Dimensions in the Customized/Adapted category generally represented categories that had been uncovered in research on psychological features of situations (e.g., Block & Block, 1981; Reis, 2008; Yang, Read, & Miller, 2006), but were not generally included in references specific to organizational contexts.

Fifth, having collected and reviewed the aforementioned measures given the considerations discussed in the previous paragraph, items were adapted from or created based on the sources discussed above, particularly the O*NET Work Context 56-item scale, the Riverside Situational Q-Sort v. 3.15, the cues listed in Tett and Burnett's (2003) Table 2, and Morgeson and Humphrey (2006). After compiling all descriptors and items from these sources, items that would be obviously irrelevant for present purposes were removed. For example, several RSQ items were removed because they pertain to situational features not likely to be found in an SJT

administered in organizational or educational contexts (e.g., "Affords an opportunity to ruminate, daydream or fantasize.") or because there was no obvious theoretical relevance associated with the items (e.g., Affords an opportunity to express femininity.).

The final 43 items retained for inclusion in the SCI are shown in Table 7. Respondents were asked to rate the extent to which each statement in the SCI was relevant to describing each situation. Because the item content contained in the SCI is drawn from multiple theoretical domains and because there are no known studies examining a broad sampling of characteristics underlying SJT item stem content, there is no known dimensional structure that underlies item scores that would be assumed a priori. Concurrently, however, there exists a need to reduce the information provided from the SCI down to a more tractable number of dimensions in order to test the study hypotheses of interest. Given the number of SCI items relative to the number of situations serving as the unit of analysis, items were clustered on a rational basis for composite formation. The process of composite formation and the final set of composites is discussed below in the Results section.

In addition to obtaining ratings of the psychological properties of item stems, ratings were collected on the individual behavioral response options concerning the extent to which each behavior in question reflects each of the FFM personality characteristics. Based on prior research by Motowidlo and colleagues (i.e., Kell et al., 2010; Motowidlo & Beier, 2010), the 10-item measure from Gosling, Rentfrow, & Swann (2003) was adapted for present purposes (Table 8). Ideally, information on each response option would be collected at a finer-grained level than the broad FFM characteristics. However, given that ratings must be provided on each response option, and given that item stems generally range from three to twelve response options, demands on raters quickly become problematic when using longer scales.

Respondent Data

College Board Situational Judgment Inventory (CB-SJI). The College Board Situational Judgment Inventory (CB-SJI) was developed for use in college admissions as a complement to standard admissions methods such as high school GPA and SAT/ACT. In its current form, the CB-SJI contains 36 items that pertain to situations relevant to the undergraduate context pertaining to both the task (e.g., being in lectures, studying course material) and interpersonal (e.g., interacting with others in project teams) domains. Each item stem is associated with five response options. The CB-SJI is administered with instructions for respondents to choose both a most likely and least likely response. These responses were scored against an expert key developed on the basis of effectiveness judgments provided by SMEs (advanced undergraduate students). Thus, the CB-SJI was scored in the same manner as that used by Motowidlo et al. (1990), Motowidlo & Tippins (1993), and others, resulting in stem scores that range from -2 to +2. The number of cases available for the 36-item CB-SJI is approximately 3,800; a subset of that sample (approximately 640) also has data for an extended 57-item form of the CB-SJI. Approximately 530 of the respondents took the CB-SJI during the college admissions process; the remaining 3,300 respondents took the CB-SJI as college students.

Criterion measures included in the CB-SJI dataset include the following: yearly GPA, four-year composite GPA, self-rated behaviorally anchored rating scales (BARS) on various dimensions of undergraduate academic performance, satisfaction (academic, social), and organizational citizenship behavior. Individual difference measures included in the CB-SJI dataset include the following: personality in terms of the FFM personality characteristics measured using the International Personality Item Pool (IPIP; Goldberg, 1999; Goldberg,

Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006) and cognitive ability measured using SAT/ACT.

Managerial Situational Judgment Inventory (M-SJI). The M-SJI is a situational judgment test developed for selecting entry-level managerial personnel. Development of M-SJI content was based on the Borman and Brush (1993) taxonomy of managerial behavior. Similar to the CB-SJI, the M-SJI was administered with instructions for respondents to choose both a most likely and least likely response. These responses were scored against an expert key developed on the basis of effectiveness judgments provided by SMEs (30 management and industrial/organizational psychology graduate students). Also similar to the CB-SJI, the M-SJI was scored in the same manner as that used by Motowidlo et al. (1990), Motowidlo & Tippins (1993), and others, resulting in stem scores that range from -2 to +2. In total, the M-SJI development item bank includes 174 items. A random selection of these items will be used for the present analyses. Respondents were undergraduate Business and Psychology students; item-level sample sizes ranged from 251 to 268.

Individual difference measures included in the M-SJI dataset include the following: personality in terms of the FFM personality characteristics measured by the NEO-FFI (Costa & McRae, 1992), cognitive ability measured using the ACT, and experience as measured based on both the number of business courses one has taken as well as the number of years of working experience one has.

Team role test (TRT). Mumford et al.(2008) introduced the Team Role Test (TRT) as an SJT developed to measure knowledge of ten team roles derived from a review of the teams and small groups literature (see also Mumford et al., 2006). The TRT contains ten item stems and ten response options for each item stem reflecting different team roles. Each item stem is associated

with a unique team role, in that the demands presented in the stem are best resolved by the individual knowing to assume the associated role. The TRT was administered with instructions to rate the effectiveness of each of the ten response options for a given item stem on five-point Likert-type scales, with higher ratings indicating greater perceived effectiveness. Thus, unlike the CB-SJI and the M-SJI, every response option for the TRT is rated by each respondent. Responses were scored in a rational/theoretical manner; endorsement of role-inconsistent behavior for a given situation results in the respondent receiving a lower score than does endorsement of role-consistent behavior.

The sample included in the TRT dataset included approximately 570 undergraduate and graduate students enrolled in management courses at a large Midwestern university. The TRT dataset was used for the purposes of estimate response option correlations with other individual difference characteristics. Individual difference measures included in the TRT dataset include the following: personality in terms of the FFM personality characteristics measured by the NEO-FFI (Costa & McRae, 1992), trait positive and negative affectivity (Watson & Clark, 1994; Watson, Clark, & Tellegen, 1988), and cognitive ability measured using the Wonderlic (Wonderlic, 1999).

Pilot Study: Situational Characteristic Ratings and Behavioral Characteristic Ratings

A pilot study was undertaken upon finalization of the content for the situational characteristic. The goals of the study were twofold. The first goal was to ensure that the instructions provided to respondents were comprehensible and capable of being followed. The second goal was to estimate the number of raters required to obtain reliable measurements of the characteristics under study. The pilot study was undertaken in the manner described below.

First, five items were randomly selected from the CB-SJI to administer to pilot study participants. Items were selected by drawing random numbers using the pseudo-random number generation procedure in Microsoft Excel 2007. Each item selected from the CB-SJI includes an item stem describing the critical incident as well as five response options. Second, having selected these five items, survey content was developed in Microsoft Word 2007 for both the situational characteristic and behavioral characteristic ratings. The survey content administered to pilot study participants is presented in Appendices C and D, respectively. Third, participants were sampled using a convenience sampling method. Participants all had at least a Bachelor of Arts or Bachelor of Science degree, and represented an array of professional backgrounds (psychology, education, fine arts, plant biology). Fourth, survey content was administered to participants via e-mail. Finally, survey responses were collected from participants, again via e-mail. Separate datasets of survey responses were created for the situational characteristic and behavioral characteristic ratings.

After collecting the respondents' ratings, analyses were conducted to ascertain similarity in ratings across raters so as to justify aggregation and to determine the number of raters required to yield adequate reliability and agreement. Intraclass correlation coefficients (ICCs) index similarity in terms of both absolute and relative agreement in ratings (LeBreton , Burgess, Kaiser, Atchley, & James, 2003; LeBreton & Senter, 2008). Thus, ICCs were calculated for each of the items. ICCs range from 0.0 to 1.0, with values closer to 1.0 indicating greater similarity. Separate ICC estimates exist for indexing rater consistency versus absolute agreement and consistency (McGraw & Wong, 1996). Both consistency and absolute agreement are necessary considerations when the purpose of ICC estimation is to justify aggregation (LeBreton & Senter, 2008); thus, ICCs assessing absolute agreement and consistency were estimated. Given

differences across (and within) literatures in notation associated with ICCs, the term ρ will be applied generically for all cases (e.g., Wong & McGraw, 1999; Zhou, Muellerleile, Ingram, & Wong, 2011), keeping in mind that this will be used to refer solely to the absolute agreement index. Because the structure of the data for the situational characteristic ratings was somewhat different from that for the behavioral characteristic ratings, slightly different models were used in the calculation and interpretation of ICCs. Each is described below.

Situational characteristic ratings. Situational characteristic ratings were collected from three raters. ICCs for the situational characteristics were estimated using a two-factor crossed random-effects model using the lme4 package in R (Bates, Maechler, & Bolker, 2011). Rater (*j*) and item stem (*s*) were modeled as crossed factors, denoted *s* x *j*, as all raters judged all item stems on each of the 43 situational characteristics. This model corresponds to Shrout and Fleiss' (1979) case 2, yielding ICCs indexed by Shrout and Fleiss as ICC(2,1) for the single-rater instance and ICC(2,*k*) for estimated ICC(2) values for *k* raters. The computation of ρ based on Shrout and Fleiss' ICC(2,1) provides a ratio of the estimated between-situations rating variance to total ratings variance, where total ratings variance includes variance due to raters, rater-by-situation interactions, and residual variance. Calculation of ICC(2,*k*) in a two-factor crossed model entails application of the Spearman-Brown prophecy formula to step up the ICC(2,1) estimate to the desired number of raters (Brennan, 2001). ICC(2,1) and ICC(2,*k*) values were estimated for each of the 43 situational characteristic ratings; ICC(2,*k*) values were estimated for number of raters ranging from two to 20.

Table 9 provides descriptive statistics for the ρ values for the 43 situational characteristic ratings; Figure 1(a) shows the mean ρ estimate plotted at values of *k* ranging from two to 20. The mean and standard deviation of the single-rater ρ values across situational characteristics were

0.37 and 0.30, respectively. However, examination of the individual ρ values for the single-rater case, ICC(2,1), for the 43 situational characteristics revealed the existence of three characteristics estimates equal to 0.00 (items 19, 37, and 40). Closer inspection of the item statistics suggested that the items in question tended to exhibit not disagreement across raters, but restriction across situations. For instance, one item (item 40) exhibited observed *SD*s of 0.58 for stem 1 (*min* = 4, *max* = 5), 1.00 for stem 2 (*min* = 3, *max* = 5), 0.58 for stem 3 (*min* = 4, *max* = 5), 0.58 for stem 4 (*min* = 4, *max* = 5), and 0.58 for stem 5 (*min* = 4, *max* =5). Such standard deviation, minimum, and maximum values suggest that raters tended to agree in their ratings of the situations for these characteristics.

As an index of inter-rater agreement, r_{wg} is less susceptible to range effects than are ICCs (LeBreton et al., 2003)[2]. Therefore, the single-item r_{wg} was calculated to further understand rater agreement for the SCI items that appeared problematic in terms of the intraclass correlations discussed above. Across all 215 ratings (43 characteristics * five item stems), the mean r_{wg} value was 0.63 (*SD* = 0.20). Concerning the three SCI items mentioned above with low observed ICC(2,1) estimates, the r_{wg} estimates averaged across the five item stems were 0.70, 0.27, and 0.77 for SCI items 19, 37, and 40. The mean r_{wg} estimates for items 19 and 40 are satisfactory. However, the mean estimate for item 37 ("P is the focus of attention or is being

[2] The accuracy of r_{wg} as an index of agreement is, however, adversely affected by the number of raters, *k* (Brown & Hauenstein, 2005; Lindell & Brandt, 1999); as *k* decreases to the number of raters used in the pilot study reported herein, r_{wg} yields underestimates of rater agreement (Kozlowski & Hattrup, 1993). Thus, r_{wg} is used solely to supplement intraclass correlations for items that appear problematic, and is interpreted in a cautious manner.

evaluated.") is quite low, suggesting that the three raters did not agree in their ratings for this item on average. From a theoretical perspective, this item would seem useful in that it appears to differentiate situations where one is being observed or evaluated by others, which may affect the types of behaviors that would be deemed appropriate (some behaviors might seem less appropriate in situations where it is known that one's behavior is being observed or evaluated). It is possible that the wording of the item is too ambiguous with regard to which party is focusing its attention or evaluating P in the situation (e.g., comparable peers, others who might have greater power or status over P and who may have authority to give rewards or social sanctions), and that the ratings provided by raters will vary depending on what raters infer in this regard.

Concerning the original point, the low single-rater ρ values for two of the three SCI items appear to indicate restriction in range across situations as opposed to lack of agreement for two of the three items in question. Although seemingly paradoxical, previous researchers have noted similar findings concerning indices of rating similarity in the presence of restriction of range (LeBreton et al., 2003). It is possible that the five SJT items sampled for the pilot study did not vary sufficiently in terms of the characteristics measured by the items that yielded single-rater ρ values of zero (e.g., because of a failure in the random sampling process used to select item stems). In any case, given that the single-rater ρ values of zero discussed above did not suggest lack of rater agreement, descriptive statistics were recomputed based on the remaining 40 items. Results are shown in Table 10; Figure 1(b) shows the recomputed mean ρ estimate plotted at values of *k* ranging from two to 20. According to Table 10, ten raters would be sufficient for reaching a mean ρ estimate of 0.70; beyond ten raters, increments in ρ values are quite small.

Behavioral characteristic ratings. Behavioral characteristic ratings were collected from three different participants from those who provided situational characteristic ratings. ICCs for

the behavioral characteristics were estimated using a three-factor nested random-effects model. Behavioral characteristic ratings were made on each of the individual response options, which can be viewed as nested within item stems. That is, each item stem is associated with five response options, but the response options differ for each stem. All raters judged all response options on each of the ten behavioral characteristics. Thus, the structure of the behavioral characteristics data corresponds to what would be denoted as (*o* : *s*) x *j*, with the terms in parentheses indicating that response options, *o*, are nested within stems, *s*, which are observed by all raters, *j*.

Variance in behavioral characteristic ratings was decomposed into components associated with the rater and stem main effects, the rater-stem interaction terms, and the response option effect. As was the case for the situational characteristics, estimates of ρ are provided for both the single- and multiple-rater cases, with number of raters again ranging from two to 20. The Spearman-Brown formula cannot be applied to a model with three factors (Brennan, 2001); thus, a modification of Wong and McGraw's (1999; equation 3.2.1) prophecy formula for a three-factor nested design applicable to ICCs indexing absolute agreement was applied in order to estimate composite reliability following aggregation across raters. Specifically, denoting the number of raters as n_k and with each rater being administered five item stems, the ICC for *k* raters was computed as:

$$\frac{\widehat{\sigma_o^2}}{\widehat{\sigma_o^2} + \frac{\widehat{\sigma_j^2}}{n_k} + \frac{\widehat{\sigma_{js}^2}}{n_k} + \frac{\widehat{\sigma_s^2}}{5} + \frac{\widehat{\sigma_{res}^2}}{(5n_k)}} \quad (1)$$

where $\widehat{\sigma_o^2}$ is the estimated variance across response options, $\widehat{\sigma_J^2}$ is the estimated variance across raters, $\widehat{\sigma_{JS}^2}$ is the estimated variance due to the rater-stem interaction, $\widehat{\sigma_s^2}$ is the estimated variance due to stems, and $\widehat{\sigma_{res}^2}$ is the estimated residual variance term.

Table 11 provides descriptive statistics of the ρ values for the 10 behavioral characteristic ratings. Figure 2 shows the mean ρ estimate plotted at values of *k* ranging from two to 20. The mean and standard deviation of the single-rater ρ value across behavioral characteristics were 0.16 and 0.14, respectively. Estimated ρ reached 0.70 at 16 or more raters. Examination of the ρ values for the ten behavioral characteristic items revealed one characteristic with a single-rater ρ of 0.00, namely the reverse-coded Openness to Experience item. Excluding this item, the mean single-rater ρ was 0.17, and ρ values in excess of 0.70 were observed when the single-rater estimate was stepped up to eight or more raters.

To corroborate the findings for the reverse-coded Openness item, r_{wg} was computed for the behavioral characteristic items (across all ten items, *M* = 0.53, *SD* = 0.37). In accord with the low intraclass correlation estimate reported for the reverse-coded Openness item above, the corresponding r_{wg} value for this item, averaged across response options, was 0.32 (*SD* = 0.34 across the 25 response options). Although the values for r_{wg} appear somewhat low for the behavioral characteristic items, the findings are difficult to interpret in light of prior research using this scale for similar purposes (e.g., Kell et al., 2010; Motowidlo & Beier, 2010), given that these researchers reported only indices of rater reliability (both the Kell et al. [2010] and Motowidlo & Beier [2010] studies reported Cronbach's α, apparently using raters as items). In

both the Kell et al. (2010) and Motowidlo & Beier (2010) studies, estimates of inter-rater reliability were quite high; three separate samples comprising six to seven undergraduate and doctoral student raters yielded reliability estimates 0.88 to 0.95. Because the remaining four FFM characteristics are measured using two separate items that sample adjectives associated with the positive and negative extremes of the trait dimensions, it would obviously be inconsistent to assess Openness using only a single item. Given this, as well as the prior findings cited above and the lack of anything aberrant or questionable with the adjectives used for the reverse-coded Openness item that strikes one as being questionable, the item was kept in its current condition.

Summary of pilot study results. Two primary conclusions can be drawn from the pilot study results described above. First, the stepped-up ICC values estimated for both the situational characteristic ratings and the behavioral characteristic ratings suggest that mean ρ values approach 0.70 as the number of judges approaches ten per rating. This number affords not only sufficient average reliability and agreement, but is also practically feasible in terms of sampling demands. Assume that each rater was asked to rate five SJT item stems, again so as to keep the number of total ratings low enough to not be burdensome or fatiguing for raters. If situational characteristic ratings had to be collected for 100 SJT item stems, then 200 raters would be required (20 blocks of five stems apiece, each block being judged by ten raters). Second, although there were several instances of situational and behavioral characteristics with very low ICC and r_{wg} estimates, the overall results suggest that individuals are capable of attaining satisfactory agreement in rating SJT item content using these characteristics. This is particularly notable for the situational characteristic inventory items, which have not yet been applied to the evaluation of SJT item stems.

With the pilot study results finalized, the remaining steps in data collection and analysis can be summarized in brief as follows:

1. Data Collection
    a. Collect ratings of 43 situational characteristics for item stems in the CB-SJI, M-SJI, and TRT from undergraduate students. Approximate number of item stems will be 90-100. Ratings for item stems were made in blocks of five stems per rater; ratings were made on each item stem by ten raters.
    b. Collect ratings of ten behavioral characteristics for response options associated with the item stems in the CB-SJI, M-SJI, and TRT from undergraduate students. Number of response options varies per test (five to ten). Ratings for response options were made in blocks of five to ten stems (and associated response options) per rater; ratings were made on the response options associated with each item stem by ten raters.
2. Estimate inter-rater reliability and agreement and between-unit variability (Research Questions #1 and 2).
3. Form composites for the situational characteristics ratings on a rational basis.
4. Conduct any required data preparation (e.g., merging, cleaning) on the respondent datasets for the CB-SJI, M-SJI, and TRT.
5. Compute zero-order correlations between SJT response data and (a) individual difference characteristics and (b) criterion measures in the CB-SJI, M-SJI, and TRT datasets.
6. Merge situational characteristic and behavioral characteristic ratings with correlations for the CB-SJI, M-SJI, and TRT to create the stem-level and response option-level datasets.
7. Run analyses testing Hypotheses 1, 2, 3a, and 3b (see below).

8. Report results.

Analysis

Prior to aggregating the situational characteristic and behavioral characteristic ratings for analysis, the models run on the pilot study data were re-estimated for the data collected from participants in the actual study to ensure adequate inter-rater reliability and agreement to justify aggregation. Individual situational characteristic or behavioral characteristic items that do not appear to function correctly (e.g., yield low ICC values in conjunction with displaying poor convergence across raters or do not appear to distinguish item stems) were excluded from analyses associated with the primary hypotheses of interest.

Stem and response option ratings were then averaged across raters to compute a single score for each characteristic for each situation and for each response option (e.g., as shown in Tables 2 and 4). These averaged ratings were merged with other data to create two datasets that were used for further analysis. The first dataset, which looks similar to Table 2, will be referred to as the *stem-level dataset*. The stem-level dataset contained average stem characteristic ratings, zero-order correlations between stem scores and individual difference characteristics, and zero-order correlations between stem scores and criteria. The second dataset, which looks similar to Table 4, will be referred to as the *response option-level dataset*. The response option-level dataset contained average stem characteristic ratings, average response option characteristic ratings, zero-order correlations between response option scores and individual difference characteristics, and zero-order correlations between response option scores and criteria.

Using the stem-level dataset, Hypothesis 1 was tested by computing zero-order correlations between stem characteristics and correlations between stem scores and other

individual characteristics. In addition, correlations between stem scores and other individual characteristics were regressed onto the set of stem characteristics in order to examine the joint effects of stem characteristics when considered as a set as well as the amount of variability explained by the model. Also using the stem-level dataset, Hypothesis 2 was tested by computing zero-order correlations between stem characteristics and correlations between stem scores and criterion measures. In addition, correlations between stem scores and criterion measures were regressed on the set of stem characteristics in order to examine the joint effects of stem characteristics when considered as a set as well as the amount of variability explained by the model.

Hypotheses 3a and 3b were tested using the response option-level dataset. Response options are treated as nested within item stems; thus, mixed-effects modeling was used to test Hypotheses 3a and 3b. Justification for and explanation of these models is elaborated upon in greater detail in the Results section. In brief, using the levels nomenclature common in discussions of multilevel modeling in the context of organizational research, response options represent level-1 units nested within item stems, which correspond to level-2 units. Trait expression ratings associated with the response options are thus level-1 predictors, whereas situational characteristic ratings associated with the item stems are level-2 predictors. The level-1 outcomes are the correlations (or their corresponding *z*-scores following transformation via the Fisher *r*-to-*z* transformation in order to normalize the correlations) between response option scores and either individual difference characteristics or criterion measures. Hypotheses 3a and 3b were tested by regressing the level-1 correlations on (a) level-1 trait expression ratings associated with response options; (b) level-2 situational characteristic ratings associated with item stems, and (c) the interactions between the predictors in (a) and (b). Statistical inference

associated with Hypotheses 3a and 3b pertains to the cross-level interaction terms. Analyses associated with the study hypotheses were conducted in R v. 2.14.2–2.15.2 (R Development Core Team, 2012).

RESULTS AND DISCUSSION

Discussion of the study results is divided into three sections. Following each section is a brief discussion and summary of the results relevant to that section. The first section, entitled *Individual-Level Characteristics of the Situational and Behavioral Characteristic Ratings*, addresses Research Questions 1 and 2, describing statistical and psychometric characteristics of the situational and behavioral characteristic ratings at the individual rater level of analysis and examining variance components to estimate between-stem and between-option variability in situational and behavioral characteristics, respectively. This section focuses primarily on the examination of consistency and agreement among raters, with the intention being to evaluate the suitability of the ratings for aggregation to the stem and response option levels for subsequent analysis. This section largely mirrors the discussion in the Method section pertaining to the rater agreement and consistency results obtained in the pilot sample.

The second section, entitled *Item Stem and Response Option-Level Characteristics of the Situational and Behavioral Characteristic Ratings and Psychometric Outcomes*, describes features of the behavioral and situational characteristic ratings at the response option- and stem-levels of analysis. This section is included to describe and illustrate the data at the level at which they will be analyzed in the third section where the substantive hypotheses are addressed. Hence, having considered issues concerning suitability for aggregation in the first subsection, this subsection describes statistical properties associated with the aggregated ratings (e.g., central tendency and dispersion, intercorrelations, composite reliability). Also examined in this section are distributional characteristics associated with the outcomes being modeled; that is, response option- and stem-level zero-order correlations between test responses and other variables of interest.

The third section, entitled *Tests of Focal Hypotheses*, describes results associated with the hypotheses posited in the Introduction. Again, the underlying purpose of the present study is to ascertain the feasibility of modeling response option- and stem-level correlations between test responses and other variables of interest by incorporating information regarding the stems and response options contained within the test. The third section, broken into two subsections, discusses results obtained from models estimated to address this issue. The first section, entitled *Results: Hypotheses 1 and 2*, focuses on analyses at the stem level; the second section, entitled *Results: Hypothesis 3a and 3b*, focuses on analyses and the response option level.

Individual-Level Characteristics of the Situational and Behavioral Characteristic Ratings

Situational Characteristic Ratings

Prior to analysis, situational characteristic composite scales were formed. The intention of composite formation was to reduce the number of predictor variables in the models while retaining substantively meaningful dimensions of situational attributes; that is, composites which retain thematically important aspects of situations based on rational and theoretical grounds. Grouping of items into scales was done on a rational basis, the process of which is as follows. First, given the number of items, the diversity of the item content, and the need for a relatively small number of scales for interpretative and estimation purposes, it was determined that seven to nine clusters of items would serve as a useful target. Eight groupings of items resulted from an initial sorting into clusters that were homogeneous or internally consistent with respect to thematic content.

After items were sorted into clusters, names and brief descriptions were attached to the eight clusters. The item content as well as the names and descriptions of the clusters were

provided to a Ph.D-level, experienced industrial/organizational psychologist. This individual placed the items into clusters to ascertain the reproducibility of the original clusters. Instances of disagreement between the two sets of clusters were discussed, with the result being the reallocation of several items to different clusters and six items were dropped. The resultant eight clusters were labeled as follows: (1) Task Demands, (2) Competition and Power (3) Interpersonal Relations, (4) Moral Issues and Fairness, (5) Individual-Emotional, (6) Familiarity and Difficulty, (7) Social Pressures and Performance, and (8) Team Task Work. Descriptions and items associated with each of these clusters are shown in Table 12.

Having sorted the items into clusters, the next step was to examine the two research questions posited in the Introduction with respect to the situational characteristic ratings. Research Question #1 pertained to the reliability and agreement of the situational characteristic ratings. To address RQ #1 pertaining to agreement in ratings, inter-rater reliability and agreement were assessed via intraclass correlation coefficients (ICCs) estimated using the model discussed in the Method section for the situational characteristic items. To review briefly, the model suggests that rating variance can be decomposed into components associated with raters, SJT stems, and error. Estimates for these variance components are then used to compute ICCs, alternatively designated as ρ_{abs} by Wong & McGraw (1999, p. 273). Estimates associated with ICC(2, 1) convey the reliability of a single rater's evaluation for any randomly selected rater; estimates associated with ICC(2, *k*) pertain to the mean rating pooled across raters as a function of the number of raters, *k*¸ associated with each rating (e.g., McGraw & Wong, 1996; Shrout & Fleiss, 1979). An ICC was analogously estimated for the behavioral characteristic ratings based on a model that decomposes rating variance into five components: raters, response options, stems, the rater-stem interaction, and error.

Because aggregation of ratings across raters for a given target is contingent upon agreement in raters' evaluations and because ICC(2, *k*) is sensitive to both absolute agreement and consistency in rank order of situations across raters, this estimate was used to justify aggregation from the individual rater level to the stem level of analysis (LeBreton & Senter, 2008). Large values indicate both agreement and consistency; low values can be attributed to lack of agreement, lack of consistency, or both. As a consequence of study design decisions to combat rater fatigue and boredom, different groups of raters evaluated different "batches" of item stems on the situational characteristics items. Because the number of raters varied across stems, a procedure outlined by Putka, Le, McCloy, and Diaz (2008) was used to estimate ICC(2, *k*).

Table 13 provides descriptive statistics for the 43 situational characteristic ratings at the individual rater level as well as estimates for the intraclass correlation coefficients, ICC(2, 1) and ICC(2, *k*). Across the 43 characteristics, the mean ICC(2, 1) value was .183 (*SD* = .080); the mean ICC(2, *k*) estimate was .661 (*SD* = .134). Values for ICC(2, *k*) ranged from .315 for item 40 to .832 to item 10. Table 14 provides descriptive statistics for the situational characteristic composite scores at the individual rater level as well as estimates for the intraclass correlation coefficient, ICC(2, 1) and ICC(2, *k*). Across the eight scales, the mean ICC(2, 1) value was .227 (*SD* = .108); the mean ICC(2, *k*) estimate was .710 (*SD* = .153). Values for ICC(2, *k*) ranged from .394 to .880 across the eight scales.

By way of comparison, LeBreton and Senter (2008, p. 839) suggested that researchers might opt for ICC(2, *k*) values ranging between .70 and .85 to justify aggregation across raters for well-established measures. Estimates for ICC(2, *k*) were equal to or greater than .70 for 22 of the 42 situational characteristic items, or roughly one-half (51.2%) of all items. Estimates for

ICC(2, *k*) were equal to or greater than .70 for six of the eight situational characteristic scales. Two of the scales had ICC(2, *k*) values less than .70: Individual Emotional, .585; Familiarity & Difficulty, .394. Given that neither of these scales has been utilized in prior research, they were retained in order to ascertain whether they were useful in modeling response option- and stem-level correlations.

RQ #2 pertained to between-unit variance in scores on the situational characteristic ratings; that is, whether the ratings sufficiently distinguish among item stems to justify using them as predictors. To address RQ #2 pertaining to between-stem variability in ratings for each situational characteristic rating, comparisons were conducted between two models: a restricted model that assumed that all variance in ratings constituted either between-rater variance or error (i.e., invariance across item stems) versus a model that permitted between-stem variance. Separate models were estimated for both the 43 situational characteristic items as well as the eight situational characteristic scales. Information criterion (AIC, BIC) and deviance (-2LL) indices were used to compare the two models for each rating. Inference was carried out using the likelihood ratio test for nested model comparison. Specifically, twice the difference in the unsigned -2LL estimate between the two models is distributed is a χ^2-variate, in this case based on 1 degree of freedom arising from estimation of the single variance parameter. Smaller AIC and BIC estimates for a target model relative to some simpler referent model and a significant χ^2 suggest that the increased complexity associated with the target model, relative to the referent, is justified on the basis of improvement in model fit.

Table 15 provides model statistics for the 43 situational characteristic items; Table 16 provides model statistics for the eight situational characteristic composites. In each table, the first

columns provide information criterion and -2LL estimates for the model forcing ratings to be fixed over stems. The second set of three columns provides information criterion and -2LL estimates for the model permitting ratings to vary over stems. The final two columns show the χ^2 estimate and *p* value based on one degree of freedom. Concerning the situational characteristic items (Table 15), permitting between-stem variance in ratings demonstrates a significant improvement in model fit. Specifically, all χ^2 estimates were significant at the $p \leq .001$ level, and the AIC and BIC estimates were uniformly smaller for the model permitting between-stem variance in ratings.

Similar results were observed for the situational characteristic composites (Table 16). All χ^2 estimates were significant at the $p < .001$ and the AIC and BIC estimates were smaller for the between-stem variance models relative to the models that forced the between-stem variance estimate to zero. To further illustrate, Table 14 provides variance component estimates associated with item stems for each situational characteristic composite (analogous estimates or the situational characteristic items can be found in Table 13). The residual variance component, which confounds error and effects associated with any rater-stem interaction given the present design, is the largest component for all eight situational characteristic scales. For some of the scales, item stems contribute equivalent or greater amounts of variation in ratings relative to raters (Task Demands, Morality & Fairness, Team Task Work), whereas variability in ratings associated with raters appears to be greater than variability associated with stems for other scales (e.g., Competition, Interpersonal Relations, Individual Emotional). In no cases, however, did the stem variance component approach zero and, as mentioned above, removing between-stem

variability from the model resulted in a significant detriment to model fit. These results provide an affirmative response to RQ #2 for the situational characteristic ratings.

Behavioral Characteristic (FFM Trait Expression) Ratings

Table 17 provides descriptive statistics for the 10 behavioral characteristic ratings associated with FFM trait expression at the individual rater level as well as estimates for the intraclass correlation coefficients. Across the ten characteristics, the mean ICC(3, 1) value was .189 (*SD* = .056); the mean ICC(3, *k*) estimate was .904 (*SD* = .053). Values for ICC(3, *k*) ranged from .783 for the negatively-worded Openness item to .958 for the positively-worded Conscientiousness item. Table 18 provides descriptive statistics for the behavioral characteristic composite scores at the individual rater level as well as estimates for the intraclass correlation coefficients. Across the five scales, the mean ICC(3, 1) value was .243 (*SD* = .052); the mean ICC(3, *k*) estimate was .939 (*SD* = .020). Values for ICC(3, *k*) ranged from .911 to .963 across the five scales. By way of comparison, LeBreton and Senter (2008, p. 839) suggested that researchers might opt for ICC values ranging between .70 and .85 to justify aggregation across raters for well-established measures. Given these results, the five FFM trait expression variables were retained as response option-level predictors.

Analogous to the situational characteristic ratings, model comparisons were conducted to ascertain whether there exists sufficient between-stem variability to justify aggregation over raters. Table 19 provides model statistics for the ten behavioral characteristic items; Table 20 provides model statistics for the five behavioral characteristic composites. Concerning the behavioral characteristic items (Table 19), the model permitting between-response option variance in ratings demonstrates a significant improvement in model fit as indicated by the significant χ^2 estimates (all estimates were significant at the $p \leq .001$) as well as the AIC and

BIC estimates (estimates were smaller for the model permitting between-stem variance in ratings). Similar results were observed for the behavioral characteristic composites (Table 20); all χ^2 estimates were significant at the $p < .001$ and the AIC and BIC estimates were smaller for the between-response option variance models relative to the models that forced the between-response option variance estimate to zero. These results provide an affirmative response to RQ #2 for the behavioral characteristic ratings.

Summary of Results: Research Questions #1 and 2

Prior to addressing the primary hypotheses of interest, analyses of the situational characteristic and FFM trait expression ratings were conducted to address two questions: (1) does sufficient inter-rater reliability and agreement exist to justify aggregation to the levels of interest, and (2) does sufficient rating variance exist across the units of measurement to justify using the aggregated ratings as predictors within the models. Inter-rater reliability and agreement for the eight situational characteristics ranged from .394 to .880. With the exception of the Familiarity & Difficulty and Individual Emotional scales, ICC(2, *k*) values exceeded .70 (Table 14), which is suitable for the purposes of aggregation (RQ #1). Because the Familiarity & Difficulty and Individual Emotional scales were developed for present purposes, they were retained for analysis despite the low ICC estimates. Future use of these two scales may warrant modification in procedure to address the low inter-rater reliability estimates observed in the present study (e.g., increasing number of raters for these scales, modifying the scale content, or not using these scales). In this study, evaluations of the substantive hypotheses related to these two situational characteristics should be viewed with caution.

For all situational characteristic scales, the residual variance component was the largest source of rating variance; stem and rater components varied in magnitude relative to one another.

Tests for the between-stem variance in situational ratings indicated a large decrement in fit when stem variances were fixed to zero for both the individual items and the composite scales (Tables 13 and 14), suggesting that between-stem differences are a meaningful source of variability in the ratings data (RQ #2). Larger stem variance component estimates were observed for the Task Demands (.209), Morality & Fairness (.332), and Team Task Work scales (.522), whereas the smallest estimates were observed for Individual Emotional (.046) and Familiarity & Difficulty (.060). Several potential reasons exist for the differences between scales in the amount of stem-level variance. First, characteristics such as task demands and morality may be more objective or verifiable, whereas perceptions concerning familiarity and emotional reactions may be more subject to idiosyncratic interpretation or perception on the part of the individual rater. A second related possibility is that demands pertaining to task characteristics, morality, or team work may be easier to convey in a clear, relatively unambiguous manner when the test is in a written format than are details regarding emotional reactions, familiarity, or difficulty. Thus, for the test developer confronted with the need for relatively brief item stems to meet administrative constraints, it may be easier to incorporate elements associated with task features into a scenario than it might be to capture, say, the tenseness of a situation.

With regard to the response option trait expression ratings, inter-rater reliability and agreement ranged from .911 to .963, which are adequate to justify aggregation (LeBreton & Senter, 2008). As was the case for the situational characteristic composites, residual variance was the largest source of variability in the FFM trait expression ratings (Table 18). Variance associated with differences between response options was the next largest source, with estimates ranging from .327 to .613, followed by variance due to raters, with estimates ranging from .164 to .233. Constraining the between-option variance component to zero yielded a significant

decrement in fit (Tables 17 and 18), suggesting that between-option differences are a meaningful source of ratings variability (RQ #2). Interestingly, the stem variance components were relatively small in magnitude, ranging from .000 to .083. Two of the FFM characteristics (conscientiousness and emotional stability) had estimated zero between-stem variability, suggesting little in the way of systematic between-stem differences in FFM trait expression ratings. The remaining three characteristics (agreeableness, extraversion, and openness), exhibited relatively small stem variance component estimates (.083, .041, and .030, respectively), indicating small but non-zero mean differences in ratings across stems. Finally, the stem-rater interaction variance component estimates indicated some degree of differential ranking of stems with regard to mean trait FFM scores across raters. Put another way, these estimates convey variability in rank order across raters if item stems were to be sorted in terms of mean trait expression. Agreeableness (.130) and extraversion (.127) demonstrated the largest estimates, followed by conscientiousness, emotional stability, and openness.

Item Stem and Response Option-Level Characteristics of the Situational and Behavioral Characteristic Ratings and Psychometric Outcomes

Situational Characteristic Ratings

Table 21 shows descriptive statistics, zero-order correlations, and internal consistency estimates (Cronbach's α) corresponding to the eight situational characteristic scales at the stem level of analysis (n = 90 SJT stems). Of the scales, seven had *SD* estimates near 0.40 or greater; the one exception was Individual Emotional (*SD* = 0.26). Correlations among scale scores ranged from -.404 between Competition and Familiarity-Difficulty to .686 between Moral Issues & Fairness and Interpersonal Relations. The average correlation among scores on the characteristics

was .191 (average unsigned correlation = .319). Of the 28 non-redundant correlations among the characteristics, five were equal to or greater than .500 in magnitude irrespective of sign; of these, two were greater than .600 in magnitude. Taken as a whole, the results suggest that the eight scales are moderately correlated with one another but do not appear to be redundant in terms of the information they provide about the stems. With respect to internal consistency, estimates ranged from .683 for Social Pressure and Social Performance to .922 for Team Task Work. Although the estimate for Social Pressure and Social Performance is lower than what might be desired, the values in Table 21 are generally consistent with standards concerning minimum reliability for newly developed scales used for research purposes.

Behavioral Characteristic (FFM Trait Expression) Ratings

Table 22 shows descriptive statistics, zero-order correlations, and internal consistency estimates (Cronbach's α) corresponding to the five behavioral characteristic trait expression scales at the response option level of analysis (n = 534 response options). The five scales had estimated *SD* values in the 0.70-0.85 range. Internal consistency estimates ranged from .657 for Extraversion to .883 for Conscientiousness. Observed correlations among scale scores ranged from .196 between Agreeableness and Extraversion to .765 between Agreeableness and Emotional Stability. The average correlation among scores on the characteristics was .510, which is higher than the average correlation observed among the situational characteristic scales (.319). Of the 10 non-redundant correlations among the characteristics, six were equal to or greater than .500 in magnitude irrespective of sign; of these, two were greater than .600 in magnitude. Taken as a whole, the results regarding the FFM trait expression correlations suggest that the five scales are somewhat more strongly correlated with one another than was observed with the situational characteristic scales, with correlations between two pairs of the scales (agreeableness and

emotional stability, extraversion and openness) being high enough to suggest that the scales are fairly redundant.

SJT Psychometric Outcomes

Tables 21 and 22 show descriptive statistics and zero-order correlations among the outcome variables at the stem level of analysis. As a reminder, the correlations associated with the individual stems were first transformed to the *z*-score metric using Fisher's *r*-to-*z* transformation in order to normalize the distribution of correlations for analysis. However, for the range of *r* values examined herein, the transformed values were similar to the original values (e.g., *r* values of .050, .100, .300, and .500 yield *z* values of .050, .100, .310, and .549). The variables are referred to interchangeably as correlations or *z* values going forward.

With respect to the individual difference correlations (Table 23), mean estimates ranged from .006 for Experience (Job Tenure) to .127 for Agreeableness. The average correlation among the transformed *z* values was .129 (the average correlation among the unsigned *z* was .181). Table 24 shows descriptive statistics associated with stem-level correlations with the criterion measures. The sample sizes in Table 24 vary for the criterion measures because different criteria were associated with different SJTs; namely, Deviance, OCB, and BARS were associated with the College Board SJT, whereas GPA data was available for both the College Board SJT and the Managerial SJI. Mean *z* values across item stems ranged from -.071 for Deviance: Time 2 to .118 for BARS: Time 1. The average correlation among the transformed z values was .145 (the average correlation among the unsigned *z* values was .384).

Correlations between SJT Psychometric Outcomes and Situational Characteristic Composite Scores

Table 27 shows correlations between the eight situational characteristic composite scores and associations between stem scores and the characteristic in question for the 89 stems being examined. Competition was significantly associated with stem-level correlations associated with ability (r = .301, p = .004); stems characterized as being high in competition tended to have stronger correlations with ability. Similar results were obtained for interpersonal relations and ability (r = .241, p = .023); in this case, stems characterized as placing greater emphasis on issues associated with individuals directly interacting with one another demonstrated stronger correlations with ability. Stems characterized as placing greater emphasis on issues associated with morality and fairness tended to exhibit more negative correlations with extraversion (r = -.242, p = .022). None of the correlations associated with the two experience variables was significant, which is likely to be partially attributable to the relatively small analysis n for these correlations (n = 39 stems).

Table 28 provides zero-order correlations between the situational characteristic composite scores and stem-score correlations with criterion variables. Scores on interpersonal relations and morality and fairness were significantly and negatively associated with stem-score correlations with deviance at both time points (interpersonal relations: r = -.334, p = .047 and r = -.509, p = .002 at times one and two; morality and fairness: r = -.484, p = .003 and r = -.467, p = .004 at times one and two), such that stems higher in both interpersonal relations and morality and fairness tended to have stronger, more negative correlations with deviance at both time points. Three other correlations in Table 28 were significant and positive: morality and fairness was positively associated with stem-level correlations with GPA (r = .230, p = .042), and familiarity and difficulty was positive associated with stem-level correlations with OCB at times one and two (r = .384, p = .021 and r = .351, p = .036). A number of other correlations were

moderate in magnitude, although not significant (e.g., each of competition, individual emotional, familiarity and difficulty, and team task work with deviance at time two; each of competition, interpersonal relations, morality and fairness, social pressure and performance, and team task work with OCB at time two).

Correlations between SJT Psychometric Outcomes and Behavioral Characteristic (FFM Trait Expression) Composite Scores

Tables 27 and 28 show zero-order correlations between response option trait expression scores and various individual difference characteristics and criterion variables. Given the relatively large number of response options being examined (n = 534), most of the correlations in Tables 27 and 28 are significant, although the associated standard errors are likely underestimated to some extent because they neglect potential dependence among the response options on stems. FFM trait expression scores correlated in the range of .10-.15 with response option score associations with ability. Relatively similar patterns of correlations were found for associations with both trait agreeableness and conscientiousness. In each case, conscientiousness trait expression demonstrated a fairly strong association (r = .444, p < .001 for agreeableness and r = .453, p < .001 for conscientiousness), correlations varied between the low .20s to the low .30s with response option trait expression scores for emotional stability (r = .298, p < .001 for agreeableness, r = .218, p < .001 for conscientiousness), extraversion (r = .313, p < .001 for agreeableness, r = .302, p < .001 for conscientiousness), and openness (r = .288, p < .001 for agreeableness, r = .255, p < .001 for conscientiousness), and in the .10s with trait expression scores for agreeableness (r = .198, p < .001 for agreeableness, r = .148, p = .001 for conscientiousness).

Patterns of correlations for associations with trait emotional stability and openness were similar to one another, with correlations in the mid to upper .20s with trait expression scores for conscientiousness (r = .239, p < .001 for emotional stability, r = .291 for openness) and extraversion (r = .248, p < .001 for emotional stability, r = .297, p < .001 for openness), and upper .10s to upper .20s for trait expression scores for openness (r = .198, p < .001 for emotional stability, r = .284, p < .001 for openness). Associations involving correlations for emotional stability and openness with both agreeableness and emotional stability were somewhat weaker, ranging between .05 and the mid .10s. Finally, relationships between trait expression scores and stem-score correlations associated with extraversion tended to be small in magnitude, with the only correlation above .20 being with trait expression scores for extraversion (r = .236, p < .001).

Correlations between FFM trait expression scores and response option correlations associated with the experience variables indicate that the trait expression scores were somewhat more useful in differentiating response options on the basis of relationships with experience operationalized as job tenure relative to relationships with number of business courses. All five FFM trait expression scales were more strongly associated with correlations involving job tenure than with business courses, with one of the five differences being significant (agreeableness: r = .022, p = .734 for business courses versus .170, p = .007 for job tenure; z = -2.12, p = .034; Meng, Rosenthal, & Rubin, 1992).

Table 30 shows correlations between the response option FFM trait expression scores and response option correlations involving the criterion measures. Somewhat different patterns of correlations were observed across the four sets of criteria. For deviance at both time points, trait expression conscientiousness was the stronger predictor of response option z-scores (time 1: r = .473, p < .001; time 2: r = -.470, p < .001). Thus, response options that were viewed as being

more expressive of trait conscientiousness had stronger negative correlations with deviance at both points in time. Trait expression scores for agreeableness and emotional stability were correlated with response option correlations with Deviance in the upper .10s to lower .20s (agreeableness: r = -.178, p = .015 at time 1; r = -.210, p = .004 at time 2; emotional stability: r = -.198, p = .007 at time 1; r = -.202, p = .005 at time 2). For trait expression openness, scores were significantly correlated with response option z-scores associated with deviance at time 1 (r = -.180, p = .013), but only marginally so at time 2 (r = -.142, p = .051). Extraversion was not significantly related to response option correlations with deviance at either time point (time 1: r = -.119, p = .104; time 2: r = -.020, p = .782).

Trait expression scores for all five characteristics were significantly associated with response option z-scores with GPA. Correlations associated with emotional stability (r = .197, p < .001), agreeableness (r = .208, p < .001), and conscientiousness (r = .210, p < .001) were at or near .20; correlations involving the other two characteristics were in the lower .10s (openness: r = .108, p = .025; extraversion: r = .117, p = .015). With respect to response option-level z-scores with OCB, trait expression scores for extraversion (time 1: r = .355, p < .001; time 2: r = .349, p < .001) and openness (time 1: r = .351, p < .001; time 2: r = .314, p < .001) were the strongest correlates of the FFM characteristics, ranging from the lower to mid .30, followed by conscientiousness (time 1: r = .283, p < .001; time 2: r = .221, p = .002). Trait expression scores for agreeableness were significantly related to OCB z-scores at time 1 (r = .153, p = .036), but not at time 2 (r = .083, p = .260), whereas emotional stability was not significantly related to OCB z-scores at either point in time (time 1: r = .139, p = .057; time 2: r = .046, p = .527).

Finally, with respect to response option-level z-scores with the BARS composite ratings at both time points, correlations were in the upper .40s to upper .50s for conscientiousness (time

1: $r = .585$, $p < .001$; time 2: $r = .488$, $p < .001$) and in the mid .30s to mid .40s for extraversion (time 1: $r = .428$, $p < .001$; time 2: $r = .363$, $p < .001$) and openness (time 1: $r = .458$, $p < .001$; time 2: $r = .386$, $p < .001$). Correlations ranged from the upper .10s to the upper .20s for agreeableness (time 1: $r = .266$, $p < .001$; time 2: $r = .190$, $p = .009$) and emotional stability (time 1: $r = .287$, $p < .001$; time 2: $r = .197$, $p = .007$).

Summary of Results: Item Stem- and Response Option-Level Descriptive Statistics

As a whole, scores associated with both the situational characteristic scales and the FFM trait expression behavioral characteristic scales demonstrated sufficient properties at the intended level of analysis. Concerning the situational characteristic scales, correlations among the composites were moderate in magnitude (Table 21); thus, while the characteristics are not entirely unique, they are also not so redundant as to preclude observing independent effects. Internal consistency estimates for the situational characteristic scales were generally sufficient (Table 21). Relative to the situational characteristic scales, somewhat larger intercorrelations were observed among the FFM trait expression composites (Table 22). Particularly high estimates were observed between agreeableness and emotional stability (.765) and extraversion and openness (.677), which in both cases met or exceeded the reliabilities of at least one of the scales involved. These correlations are similar in magnitude to those reported by Kell et al. (2010). In their two samples, correlations among the FFM trait expression scores ranged from .11 to .89 (Kell et al., 2010; Table 1, p. 221). In spite of the high intercorrelations, the five trait expression scales were kept distinct in subsequent analyses, given that the FFM traits are frequently measured as distinct characteristics (hence, there being precedence in terms of practice) and that there is no theoretical rationale for combining the scales.

A number of the situational characteristic scales were related to correlations with both other individual difference characteristics and various criterion outcomes. Concerning relationships between situational characteristics and correlations with other individual difference variables (Table 27), competition and interpersonal relations were both positively associated with stem-score correlations involving ability, whereas negative relationships were observed between morality & fairness and correlations with extraversion as well as individual-emotional and correlations with openness. Concerning relationships between situational characteristic scores and stem-score correlations with criterion outcomes, several of the situational characteristics were negatively related to stem-score correlations with deviance (i.e., interpersonal relations, morality and fairness, and social pressure and performance). Positive relationships between situational characteristic scores and stem-score correlations with criterion outcomes were also observed in a few instances (i.e., morality and fairness and correlations with GPA; familiarity/difficulty and correlations with OCB). Collectively, then, most of the situational characteristic scales were significantly related to correlations with either other individual characteristics or criterion outcomes. The exceptions to this statement were task demands and team task work; as shown in Tables 25 and 26, relationships involving these two scales were not significant.

Finally, there are a few conclusions relevant to the correlations involving the FFM trait expression scales. First, on the whole, relationships involving the trait expression ratings with correlations involving other individual difference characteristics and criterion outcomes (Tables 27 and 28) tended to be somewhat more consistent and stronger in magnitude than those observed for the situational characteristic scales. Of the 120 correlations involving the situational characteristic scales in Tables 25 and 26, 96 (80%) were between .00 and .20 in absolute

magnitude; conversely, of the 75 correlations involving the FFM trait expression scales, 75 (48%) were greater than .20 in magnitude.

In terms of relationships between trait expression and other individual difference characteristics (Table 29), each FFM trait expression scale was significantly related to response option-level correlations with the same FFM trait. For instance, FFM trait expression scores for openness were positively related to response option-level correlations involving openness (.284). Although this finding provides convergent evidence, Table 29 also reveals potential concerns regarding discriminant validity. Although each FFM trait expression scale was significantly related to correlations involving the same trait, each scale was also significantly related to correlations involving at least one other FFM trait, as well. Indeed, FFM trait expression scores for conscientiousness, extraversion, and openness were related to correlations with all FFM characteristics. Finally, response option-level correlations with the criterion outcomes were associated with multiple FFM trait expression scales. In addition to the correlations in Table 30 generally being significant (although again recall that the standard errors are underestimated due to potential stem dependencies), many of the relationships were also moderate to strong in magnitude (e.g., relationships between conscientiousness trait expression and deviance; relationships between extraversion trait expression and BARS). This is the first study to demonstrate that relatively sizable proportions of variability in SJT response option-level correlations with relevant criterion outcomes can be accounted for by thematic characteristics of the test content.

Tests of Focal Hypotheses

Results: Hypotheses 1 and 2

Hypotheses 1 and 2 pertain to models at the level of SJT item stems. Broadly speaking, the goal of testing these models is to explain between-stem variability in stem-score correlations with both individual difference characteristics and criterion measures using information regarding situational characteristics associated with the item stems. In other words, to what extent can knowledge about item stem situational characteristics allow us to predict the amount of information stem scores can tell us in terms of trait saturation (correlations with other individual characteristics) or criterion prediction (relationships with behavioral outcomes relevant in applied contexts)?

For analyses at the stem level, OLS estimation procedures were applied. The correlations between stem scores and the characteristic in question were regressed on the set of situational characteristics. Thus, again, the outcome being modeled was a correlation between stem scores and either other individual difference characteristics (i.e., FFM personality characteristics, experience, ability) or various criterion outcomes (e.g., BARS, OCB). For example, the model for correlations between stem scores and agreeableness (i.e., agreeableness stem-score saturation) would provide estimates for the effects of the eight situational characteristics on agreeableness stem-score saturation. These models yield three pieces of information: (1) slope estimates for each of the situational characteristics indicating the increase or decline in stem-score correlation per unit increase in the situational characteristic in question; (2) an estimated intercept, indicating the expected stem-score correlation when all situational characteristics in the model are at their mean, and; (3) an estimated residual variance, denoting between-stem variability in correlations not explained by the model. Situational characteristic scores were grand-mean centered to aid interpretation and, as previously mentioned, stem-score correlations were first transformed to the *z* metric prior to analysis.

Hypothesis 1 stated that situational characteristics will account for between-stem variability in stem-score trait saturation, that is, correlations between stem scores and other individual difference variables. Tables 29 and 30 show estimates from models where stem-level correlations were regressed on the eight situational characteristics. With respect to the model results concerning the individual difference characteristics in Table 31, multiple *R* estimates ranged from .184 for Conscientiousness to .476 for Experience (Business Courses). Thus, situational characteristics accounted for anywhere between 3-23% of the between-stem variance in correlations with the individual difference characteristics under examination. Thus, the set of situational characteristics explained non-negligible between-stem variability in trait saturation with FFM personality characteristics, ability, and experience.

The specific situational characteristics that emerged as significant predictors varied across the individual difference characteristics. Concerning extraversion trait saturation, morality and fairness was a significant predictor (b = -.045, SE = .021, r = -.242). For openness, morality and fairness was a significant predictor of trait saturation (b = -.042, SE = .017, p = .018, r = -.136). As such, stems that contained features tied to morality and fairness displayed more negative correlations with openness. None of the situational characteristics was significant in the models for agreeableness, conscientiousness, and experience (work tenure or business courses

Hypothesis 2 stated that situational characteristics will account for between-stem variability with regard to criterion-related validity. With respect to results concerning the criterion outcomes in Table 32, multiple *R* estimates ranged from .317 for GPA to .563 for deviance at time 1. Thus, situational characteristics accounted for approximately 10-32% of the between-stem variability in criterion prediction with the outcomes under examination. However, the number of significant coefficients in the models was relatively small. In particular, morality

and fairness was negatively associated with stem-score correlations with deviance at time 1 (b = -.058, SE = .024, p = .026, r = -.484), such that stems that emphasized themes or cues associated with morality, ethics, or fairness demonstrated stronger negative item-level criterion-related validities with deviance at time 1. In addition, familiarity and difficulty was positively associated with stem-score correlations with OCB at time 1 (b = .073, SE = .034, p = .042, r = .384), suggesting that stems that were perceived as being more familiar and manageable by an average person yielded scores that were more strongly related to OCB at time 1. Given the relatively small number of stems available for analysis in the stem-score models in light of the magnitude of effects observed, the lack of significant findings is perhaps unsurprising.

In summary, Hypothesis 1 was primarily supported whereas evidence concerning Hypothesis 2 was somewhat mixed. A non-negligible percentage of variance (3-32%) in stem-score correlations was accounted for by the situational characteristics across the models for both the individual difference characteristics and criterion outcomes (Tables 29-30). This finding lends support to the general idea that knowledge of situational features conveyed in SJT item stems can be used to predict the information obtained from stem scores in terms of trait saturation and criterion prediction.

Significant effects were found for at least one predictor for five of the eight models pertaining to the individual difference characteristics. However, the results also suggest that support for a specific situational characteristic varies depending on the trait or criterion in question. In particular, only a subset of the predictors in each model exhibited significant effects and the relevance of specific situational characteristics varied widely across individual difference characteristics and criterion outcomes. For instance, in the models predicting stem-level trait saturation, the interpersonal relations scale was only significant in the model for Extraversion.

Similarly, in the models predicting stem-score criterion prediction, morality and fairness was significant only in the model for deviance at time 1.

Thus, although there is general support for the relevance of situational characteristics in explaining stem-score trait saturation and criterion-related validity, the results for specific situational characteristics are generally not consistent enough to definitely state that specific characteristics are either uniquely critical or, conversely, unimportant. Potential exceptions to this statement pertain to competition (significantly associated with ability and extraversion), familiarity and difficulty (significantly associated with emotional stability, experience in terms of number of business courses, and OCB at time 1), and morality and fairness (significantly associated with extraversion, openness, and deviance at time 1). That said, caution should be exercised in over-interpreting this statement, as it is based off of patterns of statistical significance, which confound factors such as sample size, as much as practical significance in terms of the magnitude of the effect. Another potential exception pertains to the lack of significant effects for the situational characteristic scales associated with both task demands and team task work. That is, neither task demands nor team task work were significant predictors of stem-level correlations involving individual difference characteristics or criterion outcomes.

Results: Hypotheses 3a and 3b

For analyses associated with Hypotheses 3a and 3b pertaining to response option-level correlations, mixed-effects modeling was applied. As described below, mixed-effects, or mixed, models are useful in situations where variance in either response or effect is believed to be attributable to multiple sources within a given data structure. As it pertains to the present study, the rationale underlying this choice of model pertains to the structure of SJTs and the resultant dependencies that this structure would be expected to yield.

As a point of departure, SJT response options can be thought of as clustering around stems in at least two ways. First, from a design perspective, SJT developers write response options to be uniquely relevant to the referent situation described in a stem. Unless an error was made, it would be highly unlikely that one would observe the response option "Attempt to reconcile differences among the coworkers to resolve the budget debate" with a situation pertaining to an interaction with a combative, disgruntled customer in a convenience store; the response option is irrelevant and would seem bizarre in light of the situation. Thus, the situation has a very direct influence on circumscribing the potential interpretation and relevance of a response. A second and related perspective for thinking about the clustering of response options around stems pertains to the instructions provided to respondents who are being administered an SJT. An individual taking an SJT is generally prompted to read each stem to understand the problem presented by the situation. She or he is then asked to read the associated response options and evaluate them on the basis of some specified criterion (e.g., effectiveness, likelihood of doing) in light of the situation. Thus, SJT instructions prompt evaluation and judgment of response options within the context of the demands of the situation with the aim of detecting between-persons variability in such patterns of judgment, hence the name of the measurement method.

The notion of clustering in test design as described above is not unique to SJTs. Clustering of response options around stems results in a test design that consists of what have been referred to alternatively as testlets (Sireci, Thissen, & Wainer, 1991; Wainer & Kiely, 1987; Wainer & Lewis, 1990), item bundles (Rosenbaum, 1988), or context-dependent item sets (Haladyna, 1992). A prototypical example of such test designs are measures of reading comprehension, critical thinking, or other constructs in the domain of cognitive performance

common to tests such as the SAT where a series of questions follows a passage that the respondent reads and evaluates. Testlet-based structures have motivated the application of random- or mixed-effects models to account for dependencies resulting from such clustering effects (e.g., Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005). In addition to accommodating dependencies that arise from clustering, mixed models permit the researcher to examine whether contextual features associated with the clustering unit are related to observed relations occurring among the units that are clustered. Thus, as it pertains to the present study, such models can be used to estimate the effects of stem situational characteristics on associations between FFM trait expression and correlations between endorsement of that option and external variables.

To ground the aforementioned discussion more squarely in terms of the present topic, assume for the moment that a researcher is focusing on a single item stem that is associated with an arbitrary number of response options, say five. For each of the five response options, the researcher has two pieces of information in hand: (1) an estimate of the zero-order correlation between the endorsement or rating of the response option and some external variable, say a measure of task performance (i.e., a response option-level criterion-related validity), and (2) an evaluation of the level of some trait that is expressed by the behavior described in the response option, say conscientiousness. Using these data for this one stem and five response options, one could regress the validities with task performance onto the ratings of conscientiousness trait expression. Like any other regression equal, doing so would yield an equation comprising an intercept term and a slope. The slope would convey the expected increase or decrease in the response option-level validity with task performance per one-unit increase in the trait conscientiousness ratings. The intercept conveys the expected response option-level validity for

an average response option in terms of conscientiousness trait expression, assuming the trait expression ratings have been mean centered prior to analysis. This equation can be used to derive the expectation for the criterion-related validity for task performance, conditional on the parameter estimates and a score for conscientiousness trait expression.

The example in the prior paragraph pertained to a single SJT item stem. When the analysis is extended to multiple stems comprising an SJT, one has a couple of options available. On the one hand, one could apply OLS estimation procedures and ignore the possibility of clustering effects. Such an approach assumes that response options drawn from two different stems are interchangeable, and that the regression parameters are invariant across stems. In other words, the regression equation is estimated by pooling over item stems completely. Alternatively, one could entertain the possibility that clustering effects might be present, evaluate the feasibility of that argument in light of the evidence presented by the data, and then use estimation procedures that allow for clustering effects in the regression of the correlations on the response option FFM trait expression predictors. Mixed-effects models accommodate the latter approach.

Clustering effects can be inferred by the presence of non-zero between-stem variance with respect to the parameters estimated in the regressions. Conceptually, this is somewhat similar to saying that the regression intercepts or slopes estimated across response options within any given stem, *i*, may not be the same as those parameters estimated across response options within another stem, *i´*, with "not the same" implying different beyond that which might be attributable to sampling error. Thus, for each parameter, there is a presumed distribution over the stems, the variance of which is estimated from the data. The benefit of estimating variance parameters in terms of model fit can be examined for any part of the model, including the

intercept or one or any of the regression slopes. Therefore, part of the regression may assumed to be fixed or invariant across stems (e.g., the intercept or any subset of the slopes), whereas restrictions on invariance may be relaxed for other parts of the model. For parameters that are permitted to vary across stems, the stem estimates are accounted for by introducing random effects that indicate perturbations or deviations from the global estimate. The expectation of the random effects across stems is zero; variability in the random effects is captured by the aforementioned variance parameters. Although parameters may be permitted to vary across stems, it is also assumed that stems are drawn from a larger distribution characterized by the overall mean or average about which the stem estimates vary.

In addition to allowing different aspects of the regressions to vary across item stems, a second benefit of using a mixed-effects approach for present purposes, as mentioned above, is that it permits the examination of predictors associated with the item stems in addition to the predictors associated with the response options, as well as interactions between the two. Conceptually, this part of the analysis can be thought of as arising in conjunction with the aforementioned random effects and associated variance parameters. From a conceptual perspective, the stem-specific deviations associated with the random effects can themselves be modeled as outcomes that are regressed on the situational characteristic scores associated with the item stems. Modeling the random effects for the intercept terms is akin to examining whether situational characteristics associated with the item stems are related to the expected response option correlation when the FFM trait expression predictors are set to zero. Modeling the random effects for the slope terms is akin to examining whether situational characteristics associated with the item stems are related to the effect of the response-option trait expression predictors on the response-option correlations. The latter represents an interaction between the stem

characteristics and the FFM trait expression characteristics; the effect of the trait expression characteristic in predicting response option correlations depends upon the level of the situational characteristic associated with the stem.

The choice of whether to allow the intercept or slope parameters vary across stems should be informed by theory (e.g., the substantive meaning of variance in parameters given the phenomenon under study and associated theoretical considerations) as well as by what the data suggest based on considerations of empirical model fit and parsimony. Substantively, between-stem variability in intercepts suggests that stems differ with regard to the expected response option level correlation when all predictors in the model are set to zero. Between-stem variability in slope estimates suggests that stems differ with regard to the effect of some predictor on the response option correlation.

As was stated earlier and will be revisited shortly, Hypotheses 3a and 3b pertained to interactions between situational characteristics associated with the stems and FFM trait expression behavioral characteristics associated with the response options in predicting correlations. For instance, one might find that response options that are more expressive of conscientiousness are also more highly correlated with task performance on average, but the effect of conscientiousness is heightened or dampened when the referent situation emphasizes some contextual feature such as time urgency. A necessary prerequisite for the existence of such an interaction is variance in the effect of the trait expression across item stems: if there is no variability in the slope of trait expression across stems, then there can be no interaction between trait expression and a stem-level situational characteristic. Therefore, prior to modeling interactions between the FFM trait expression predictors and the situational characteristics, it should first be demonstrated that there is even a need to let the slopes associated with FFM trait

expression vary. Given the research question, the regression intercepts are of lesser interest than the regression slopes; therefore, the choice of whether to permit the intercepts to vary will be more influenced by considerations associated with model fit. Hypothesis 3a suggested the presence of interactions between item stem situational characteristics and behavioral characteristics in terms of FFM trait expression associated with the response options in predicting response option correlations with individual difference characteristics in the domains of personality, ability, and experience. Hypothesis 3b suggested the presence of interactions between item stem situational characteristics and FFM trait expression in predicting response option correlations with criterion outcomes. The following process was used in testing Hypotheses 3a and 3b.

For models predicting response option-level trait saturation (Hypothesis 3a), two approaches were taken for different subsets of individual difference characteristics: one approach for response option correlations with the FFM personality traits, and a separate approach for response option correlations with ability and experience. With regard to trait saturation for the FFM personality characteristics, one model was estimated for each of the five traits. For each FFM personality trait, only the trait expression ratings for the same trait were included as a response option-level predictor, in addition to the situational characteristic predictors. For instance, the model predicting response option correlations with agreeableness included FFM trait expression scores for agreeableness as a predictor; scores for the remaining FFM traits were not included. Similarly, for the model predicting response option correlations with conscientiousness, FFM trait expression scores for conscientiousness were included as a predictor, with scores from the remaining four FFM traits being not included.

The primary factors behind this modeling strategy were the related considerations of precision in estimation and parsimony. Concerning precision and power, even with the model structure specified as above, the full model would contain 18 fixed-effects terms: one intercept, one trait expression effect, eight situational characteristic effects, and eight interaction effects. If the model included all FFM trait expression predictors, the number of fixed-effects predictor terms would explode to 54: one intercept, five trait expression effects, eight situational characteristic effects, and 40 interaction effects. Given the sample size at the stem level for analyses associated with correlations with the FFM personality characteristics (90 stems), the estimation of 54 fixed-effects plus the variances and covariances is burdensome in precision in parameter estimation. In addition to considerations regarding power and precision, being in the position of interpreting 54 terms does not strike one as being tractable in terms of delineating reliable and meaningful patterns among the effects.

The aforementioned approach was applied to trait saturation involving the FFM personality characteristics. Concerning models associated with correlations for ability and experience, a different approach was taken. Again, the need to keep the response option-level predictors to a tractable number was present given the aforementioned concerns. However, there was also no strong a priori justification for restricting focus to any single FFM trait expression predictor when modeling trait saturation involving ability or experience; that is, it is difficult to generate a strong theoretical rationale for why the level of, say, agreeableness or extraversion associated with response options would or should be more predictive of the extent to which the response options provide information associated with ability or experience than the other FFM trait expression characteristics. Therefore, for ability and experience, a separate model was first estimated for each FFM trait expression predictor, yielding five models for each outcome. These

models provided information concerning the effect of each FFM trait expression predictor on trait saturation for ability and experience, as well as the variability in the effect across item stems. As mentioned above, variability in the slopes across stems is a necessary condition for examining interactions between the FFM trait expression and the situational characteristics. Therefore, the intent was to retain FFM trait expression predictors of ability and experience that accounted for non-zero variance in correlations between response options and demonstrated meaningful variability in slopes across item stems.

With regard to the models for the FFM personality correlations, two baseline models were fit for comparison with one another. First, an OLS model was fit whereby trait saturation for each of the FFM personality characteristics was regressed on FFM trait expression scores for the same characteristic. This model ignores the aforementioned clustering effects likely to be associated with the item stems and is thus adversely affected by model misspecification to the extent that clustering is evident (e.g., biased standard errors). The model is, however, useful for ascertaining the gain in model fit obtained by permitting the model parameter estimates to vary across stems. For the OLS models, FFM trait expression scores were grand-mean centered for analysis. Following estimation of the OLS models, mixed-effects models as described above were then estimated. For the mixed models, FFM trait expression scores were stem-mean centered prior to entry for analysis (that is, deviation scores were computed by subtracting the stem means from each rating). All mixed-effects models were estimated using restricted maximum likelihood estimation.

Table 33 provides estimated intercepts, slopes, residual *SD*s, and fit statistics for the OLS models. First, the intercepts for each model varied between .00 and -.01. Thus, for response options with near-average trait expression scores, trait saturation with the FFM personality

characteristic in question tended toward zero. A significant, positive relationship was observed between trait expression for each given FFM characteristic and saturation with the same trait: agreeableness (b = .022, SE = .005), conscientiousness (b = .043, SE = .004), emotional stability (b = .013, SE = .004), extraversion (b = .020, SE = .004), and openness (b = .031, SE = .005). Thus, response option-level trait saturation tends to increase among response options perceived as being more expressive of the trait in question. Model R^2 values varied across the five characteristics, ranging from .022 for the model for emotional stability to .205 for the model for conscientiousness.

Given the model results presented in Table 33, it is clear that response option-level trait saturation is positively related to perceived trait expression of those response options. The next issue was to examine whether the positive relationship between FFM trait expression and response option correlations with like personality traits varies across stems. As an exploratory step, estimates of the regression slopes were first plotted for each FFM characteristic. Figure 3 shows the regressions of response option correlations with the five FFM traits on like trait expression. These are OLS estimated slopes, with the slope estimated separately within each item stem with the uncentered FFM trait expression predictor. The slopes in Figure 3 suggest substantial variability between item stems in terms of FFM trait expression slopes. In addition, it appears that there is some degree of variability across the five FFM characteristics in terms of the general patterning of the slopes (e.g., slopes for conscientiousness looking somewhat more consistently positive in sign than slopes for, say, agreeableness). Although there is a general trend for the slopes in Figure 3 to be positive, it is clear that, for each of the FFM characteristics, there are stems where the regression of FFM trait saturation correlations onto trait expression

scores for the same FFM trait is negative. For these stems, as FFM trait expression shows increase, the predicted trait saturation correlation with the same FFM actually decreases.

To investigate this issue in greater detail, the five stems exhibiting the most negative slopes for each FFM characteristic in Figure 3 were examined to ascertain any patterns that might indicate an explanation for the negative slopes. One commonality shared by the stems with the most negative slopes for each of the five FFM characteristics plotted in Figure 3 was that they had below-average within-stem variability concerning the FFM trait expression ratings. That is, if one were to compute the between-option variance in the FFM trait expression ratings within each stem, those stems that had the strongest negative slopes in Figure 3 also tended to have lower-than-average variability in trait expression.

To illustrate, the five most negative slopes for agreeableness in Figure 3 were for the following items: 105 for the M-SJI (*b* = -.247), items 15 and 20 for the CB-SJI (*b* = -.233 and -.217, respectively), and 166 and 148 for the M-SJI (*b* = -.163 and -.134, respectively). The between-option *SD* in agreeableness trait expression ratings was then computed for each of the item stems, and was also converted to *z*-score metric to facilitate interpretation. Across stems, the mean between-option *SD* for agreeableness ratings was .654. The *SD* values for the five stems listed above were .437 (*z* = -.623), .441 (*z* = -.612), .113 (*z* = -1.554), .278 (*z* = -1.079), and .297 (*z* = -1.025). As indicated by the *z*-scores or by directly comparing the individual within-stem estimates against the mean estimate, the five stems had low within-stem variability in agreeableness trait expression.

For each FFM characteristic, Table 34 shows the five most negative OLS slope estimates across the 90 stems, as well as the associated *SD* of the FFM trait expression ratings across the response options within the slopes and the *z*-score corresponding to that *SD*. As was the case for

agreeableness, the *SD* values for those stems with the most negative slopes all tend to be far below that of the average or typical stem. One interpretation of low within-stem variability is that raters did not differentiate among the response options, perhaps because the behaviors were seen as similar with regard to trait expression. There are two possible explanations for why such an argument might hold.

First, it might have been the case that all options were truly similar with regard to the expression of some trait. The item stem shown below had a negative slope in the regression of conscientiousness trait saturation correlations on conscientiousness trait expression, as well as low variability across response options in ratings of conscientiousness trait expression.

You share a company account with fellow managers. You access the account to conduct company business, and discover it has been depleted.

a) Confront the managers about the situation.
b) Report it to supervisors for an investigation.
c) Contact my manager and provide her/him the receipts for all of my transactions.
d) Ask accounting for a listing of all transactions on the account to see there are excesses.

In this case, the cues in the stem suggest that conscientiousness might be involved in a response to the situation. In particular, a problem (the depleted account) is discovered in the process of carrying out a work-relevant task (conducting business) that serves as an obstacle in successfully completing the task. Furthermore, all of the response options seem to convey conscientiousness or perhaps one of its facets (e.g., all response options show initiative in that they are directly confronting the problem associating with discovering a depleted account). Thus, even though conscientiousness might be relevant to the situation and associated behaviors, the response options do not seem to vary greatly along that dimension. This point is addressed in greater detail in the Conclusion section.

A second possible reason for why ratings of trait expression may not vary among response options within a stem for some characteristic is that the situational cues in the stem or the characteristics of the behaviors in the response options may have been developed such that between-option variability in a given FFM dimension was not seen as relevant. For instance, the following stem and series of options had a negative slope between agreeableness trait expression and agreeableness trait saturation as well as low variability in agreeableness ratings across response options:

You have been working on a project to which you are highly committed for several months. Your supervisor directs you to discontinue the project.

a) Ask for specifics about why the project was discontinued and see if there is a way to continue.
b) Discontinue the project.
c) Save the work for a later time when you can return to the project.
d) Talk to your supervisor's boss about the situation

Arguably, agreeableness may not seem to be overly relevant to the situation, in that there are not obvious thematic cues, demands, or other features in the content of the stem that would implicate agreeableness. In addition, none of the response options seems overly high or low on agreeableness; rather, they seem to vary on characteristics that one might expect to be independent of agreeableness. Irrespective of which of these two explanations holds, the co-occurrence of negative slopes for stems that also exhibit low within-stem variability in trait expression seems to be a systematic finding across the five FFM characteristics.

Having examined the OLS estimates of the regressions, Table 35 shows results from a series of nested-model comparisons to ascertain which parameters in the regression equations vary across stems. The baseline model is that which permits both intercepts and slopes to vary; the other two models fix one of the two parameters. For each of the five models, a significant decrement in fit was observed when the slope was fixed across stems (comparison of model 1

versus 2), as indicated by the significant chi-square estimates and increases in the AIC and BIC values. In other words, forcing all item stems to have the same slope for the relationship between trait expression and trait saturation yielded a significant decrement in model fit relative to a model that assumed that slopes varied across stems. No decrement in model fit was observed when the intercept was fixed across stems (comparison of Models 1 versus 3), however. Thus, any observed variance in regression intercepts was sufficiently small to assume that the estimate could not be differentiated from zero. Given these results, the intercepts for the regression of FFM trait saturation correlations on FFM trait expression ratings was assumed to be fixed across stems, whereas the slopes were permitted to vary.

Table 36 shows the parameter estimates for the fixed-intercept, random-slope models. As was the case for the OLS models, the intercepts for each model varied between .00 and -.01; in addition, significant, positive relationships were observed between trait expression for each given FFM characteristic and response option correlations with that characteristic: agreeableness (b = .030, SE = .007), conscientiousness (b = .048, SE = .006), emotional stability (b = .019, SE = .006), extraversion (b = .019, SE = .006), and openness (b = .039, SE = .006). Thus, response option-level FFM trait saturation tended to be higher for those response options perceived as being higher in the characteristic in question.

Model R^2 estimates for mixed-effects models were obtained by subtracting from one the ratio of model residual variance to the observed variance (Gelman & Hill, 2009; p. 474-475). Model R^2 values again varied across the five FFM characteristics. Furthermore, the estimates were substantially larger than those observed for the OLS estimates in Table 33, ranging from .168 for agreeableness to .404 for conscientiousness (analogous estimates in Table 33 ranged from .022 to .205). Model fit, as gauged by comparison of model R^2, increased by a factor of

roughly two (.404/.205) to slightly under nine (.195/.022) depending on the FFM characteristic in question in comparing the OLS models against the mixed models. Thus, permitting the slopes of FFM trait expression to vary across item stems provides a substantial improvement in terms of modeled variance in correlations between response option scores and FFM personality characteristics. From a conceptual standpoint, this finding can be interpreted as a form of interaction between the traits expressed by the response options and the discrete situations within which those behaviors could be enacted in predicting correlations between endorsement of the behaviors and other individual difference characteristics (i.e., the extent to which the response options are saturated with these other characteristics).

Figure 4 is a density plot illustrating the distribution for the five FFM trait expression slopes across item stems from the mixed-effects models. The highly peaked distribution is for the openness trait expression slopes; given the density's location and relatively small dispersion, it is apparent that very few of the openness slopes were at or below zero. Thus, openness trait expression tended to have a positive relationship with response option-level correlations involving trait openness across the vast majority of the item stems. Although the distribution for conscientiousness trait expression slopes was more disperse than the openness distribution, there was very little overlap with zero, as well, because of its somewhat higher location. There was somewhat greater overlap between the densities for the other FFM trait expression slopes and zero.

Having estimated the overall effect of FFM trait expression on response option correlations with like traits as well as variability in that effect across item stems, the next step was to introduce the situational characteristic scales into the models. Table 36 provides model estimates for response option FFM trait saturation correlations regressed onto same-trait FFM

expression, the eight situational characteristic scales, and the interaction terms involving FFM trait expression and the situational characteristic scales. Several patterns are apparent across the models. First, as was the case for the models excluding the situational characteristic scales, the fixed-effects estimates for trait expression were significant and positive. Second, none of the situational characteristic scales had a direct effect on response option-level trait saturation in the models. Such a finding is not entirely surprising. Situational characteristics vary at the stem, as opposed to response option, level of analysis; two response options within a given stem possess the same standing for a given situational characteristic. Therefore, the main effects of situational characteristics are associated not with variability at the response option level analysis, but at the stem level of analysis. However, given the absence of between-stem variability in regression intercepts, there is no between-stem variability to be accounted for by the situational characteristics.

Of greater relevance are the interactions between situational and behavioral characteristics in the prediction of response option correlations. In three of the models shown in Table 37 (namely, agreeableness, emotional stability, and extraversion), none of the interaction terms was significant at the conventional $p < .05$ level in three of the models. Three situational characteristics demonstrated significant interactions with personality trait expression for conscientiousness in predicting response option trait saturation with conscientiousness: interpersonal relations ($b = .055$, $SE = .021$), individual-emotional ($b = -.054$, $SE = .025$), and social pressure/performance ($b = -.041$, $SE = .021$). One situational characteristic demonstrated a significant interaction with openness trait expression in predicting response option correlations with trait openness, namely familiarity/difficulty ($b = -.033$, $SE = .017$).

Figures 5-7 show plots for the estimated simple slopes pertaining to the significant and marginally significant interactions, estimated below and above one standard deviation from the mean for the moderating situational characteristic variable. The interactions graphed in Figures 5-7 are disordinal in nature. All positive interaction terms in Table 37 were associated with steeper slopes at higher levels of the situational characteristic variable; thus, the effect of FFM trait expression on response option-level same-trait saturation was stronger among stems that were higher in the situational characteristic. Conversely, all negative interaction terms were associated with steeper slopes at lower levels of the situational characteristic variable. An example of a positive interaction involved FFM trait expression conscientiousness and interpersonal relations in predicting response option-level correlations with conscientiousness (i.e., conscientiousness trait saturation). The estimated simple slope for stems low in interpersonal relations demands was $\hat{Y} = -.004 + .023X$, whereas the estimated simple slope for stems high in interpersonal relations demands was $\hat{Y} = -.007 + .074X$, with $\hat{Y}$ indicating the predicted correlation between endorsement of a response option and conscientiousness (i.e., conscientiousness trait saturation) and X indicating stem mean-centered conscientiousness trait expression scores.

An example of a negative interaction involved FFM trait expression conscientiousness and individual-emotional in predicting response option-level correlations with conscientiousness. The estimated simple slope for stems low in individual-emotional was $\hat{Y} = -.004 + .062X$, whereas the estimated simple slope for stems high in individual-emotional was $\hat{Y} = -.007 + .034X$, again with $\hat{Y}$ indicating the predicted correlation between endorsement of a response option and conscientiousness (i.e., conscientiousness trait saturation) and X indicating stem mean-centered conscientiousness trait expression scores. Thus, despite the overall positive

relationship between conscientiousness trait expression and response option scores with trait conscientiousness, the relationship is stronger among stems that emphasize certain cues (interpersonal relations, task demands) and is weaker among stems the emphasize other cues (individual-emotional, social pressure/performance, morality/fairness), thus indicating a situation-behavior interaction in predicting the amount of information response options provide about FFM trait characteristics.

Having examined response option-level correlations with FFM traits, the next step was to examine relationships between FFM trait expression and correlations involving the other two individual difference characteristics under study, ability and experience. Because the FFM traits do not map directly on ability or experience in a theoretically clean manner, the process of model development for ability and experience began by first regressing the response option-level trait saturation correlations associated with ability and experience onto trait expression for each of the FFM characteristics. Results were examined with respect to (a) the fixed-effect estimate for each FFM characteristic and (b) the variability in the FFM trait expression slopes. Starting in this manner, one can first ascertain which FFM trait expression variables seem most relevant for explaining variability between response options in correlations with ability and experience.

Table 38 shows model estimates for the regression of correlations with ability and experience on trait expression for each of the FFM characteristics. The intercept denotes the expected trait saturation correlation for a response option that is average with regard to the FFM characteristic in question. The slope corresponds to the increase in trait saturation correlation per unit increase with regard to the FFM characteristic, whereas *SD* slope provides an estimate of the standard deviation of the random effects for the FFM trait expression slopes. Relative to the models predicting FFM trait saturation, relatively little variance was explained in correlations

with ability and experience; model R^2 estimates ranging from .001 (agreeableness and experience: business courses) to .084 (openness and experience: job tenure).

Furthermore, between-stem variability in the slopes was estimated at zero across all of the models for ability and experience. Given the results observed for the models predicting correlations with FFM trait characteristics (e.g., none of the SD estimates in Table 36 was zero), this finding is somewhat surprising. In this case, a lack of variability, indicated by an estimated *SD* at or near zero, indicates that the regressions of ability and experience trait saturation onto the FFM trait expression predictors are invariant across item stems with regard to the slope estimates. Given the lack of variability in slopes across stems, there is little reason to model stem-level characteristics as predictors of slopes; there is no between-stem variability to account for with stem-level predictors. Therefore, no models were estimated with the inclusion of situational characteristics as predictors of FFM trait expression slopes for ability and experience.

Collectively, the results for the response option-level models provide mixed support for Hypothesis 3a which predicted that response option trait expression would interact with the situational cues conveyed in the item stems in predicting response option-level trait saturation (i.e., correlations between response option scores and other individual difference characteristics). No interactions were observed in the models for either agreeableness or extraversion. With regard to the models for ability and experience, between-stem variance in regression intercepts and slopes was estimated at zero, hence precluding a test of the interactive effects between response option trait expression and situational characteristics.

Hypothesis 3b stated that response option trait expression would interact with situational characteristics associated with item stems in the prediction of response option-level criterion-related validities. Similar to the analyses involving ability and experience, there was a desire to

focus on a relatively small number of response option-level predictors in the interaction models because of the potential explosion of terms in the model. Therefore, analysis began by first fitting various models to ascertain which response option-level predictors were most relevant. Relevance was ascertained by examining the magnitude of the fixed-effects estimate combined with the amount of between-stem variance in slopes. Again, if no between-stem variability in slopes exists, there is no reason for examining stem-level predictors of slopes. Having delineated which FFM trait expression characteristic(s) was most relevant, the analysis proceeded by including in the model the situational characteristics as stem-level predictors.

Tables 37 and 38 show OLS and mixed model estimates, respectively, for the regression of response option-level validities on the FFM trait expression predictors. Again, the OLS model ignores the potential for between-stem variability in intercepts or slopes, serving as a useful baseline for examining the benefit of modeling variability in slopes across stems (as was the case for the other models, estimates for between-stem variability in intercepts were zero or near zero; therefore, only variability in slopes was modeled). Concerning Table 39, only five of the 35 OLS slope estimates were non-significant, namely extraversion and deviance at times one and two, emotional stability and OCB at times one and two, and agreeableness and OCB at time two. Across all models, R^2 estimates ranged from .000 (extraversion and deviance at time two) to .343 (conscientiousness and BARS at time one). As such, the effect of trait expression on response option-level criterion-related validity varies widely across specific FFM characteristics and outcomes. Averaging the model R^2 estimates across FFM trait expression predictors for each outcome, mean R^2 estimates ranged from .030 for GPA to .178 for BARS at time one.

Table 41 provides the mixed model counterparts of the Table 39 estimates with between-stem variability in the trait expression slopes permitted. Only four of the 36 slope estimates were not significant, namely extraversion and deviance at times one and two, emotional stability and OCB at time two, and agreeableness and OCB at time two. Across all models, R^2 estimates ranged from .000 (extraversion and deviance at time two) to .425 (conscientiousness and BARS at time one), suggesting that the effect of trait expression varies widely across specific FFM characteristics and outcomes. Averaging the model R^2 estimates across FFM trait expression predictors for each outcome, mean R^2 estimates ranged from .053 for GPA to .275 for BARS at time one. The average increase in R^2 resulting from allowing the slopes to vary (i.e., Table 40 versus Table 39) relative to assuming slopes were fixed were as follows: agreeableness, .084, conscientiousness, .069, emotional stability, .133, extraversion, .042, and openness, .072. Thus, the largest increase in R^2 resulting from permitting slopes to vary was for emotional stability, where about 13% more variance in response option-level criterion-related validities was accounted for, on average, across the outcomes.

As another way of comparing the relevance of the various FFM trait expression predictors, Table 41 shows results from models where all FFM characteristics were entered together as predictors of response option-level criterion-related validities. When all five characteristics were entered simultaneously, many of the significant effects observed in Tables 37 and 38 disappear. Even still, significant FFM trait expression effects were observed in the models for criterion-related validities for each outcome. For deviance, conscientiousness (time 1: b = -.038, SE = .007; time 2: b = -.057, SE = .008) and extraversion (time 2: b = .026, SE = .009) were significant predictors. Thus, response option criterion-related validities with deviance were

more strongly negative among response options that were more expressive of trait conscientiousness. Interestingly, the opposite effect was observed for extraversion: response options more expressive of trait extraversion tended to have stronger positive correlations with deviance. With regard to GPA, the only significant predictor was agreeableness (b = .017, SE = .008), such that response option-level criterion-related validities for GPA were higher among response options that are more expressive of agreeableness.

For OCB, response option validities at time 1 were significantly and positively related to conscientiousness (b = .012, SE = .006), extraversion (b = .019, SE = .008), and openness (b = .024, SE = .012). Of these, only extraversion was significantly related to response option validities with OCB at time 2 (b = .024, SE = .010). Finally, conscientiousness (b = .046, SE = .007) and extraversion (b = .028, SE = .009) were positively related to response option validities with BARS at time 1, such that response options characterized as being expressive of conscientiousness and extraversion had stronger correlations with BARS. Only conscientiousness was significantly related to correlations with BARS at time 2 (b = .038, SE = .008). In terms of between-stem variability in the regression slopes, slopes associated with emotional stability exhibited the largest *SD*s of the FFM characteristics for deviance at times 1 and 2 (.030 and .061), OCB at times 1 and 2 (.036 and .048). Slopes associated with openness exhibited the largest SD for GPA (.023) and BARS at times 1 and 2 (.033 and.036).

Because emotional stability had the greatest between-stem variability in slopes in the single predictor models, this characteristic was retained for the models involving the situational characteristic predictors. Also, given the relatively small variance of the random effects for all predictors in the GPA models, only deviance, OCB, and BARS were examined. Figure 7 shows density plots of the emotional stability trait expression slopes for each outcome. As would be

expected, slopes for BARS and OCB tended to be located above zero (i.e., response options more expressive of emotional stability tended to have positive validities with BARS and OCB), whereas slopes for deviance tended to be below zero (response options more expressive of emotional stability tended to have stronger negative validities with deviance). The densities for the OCB stems are relatively diffuse relative to those for BARS and deviance; across stems, emotional stability slopes for OCB took on values over a much wider range than those for deviance and BARS.

Table 42 provides model estimates for the criterion-related validities associated with deviance, OCB, and BARS regressed on emotional stability trait expression, the stem-level situational characteristics, and the interaction between emotional stability trait expression and the situational characteristics. The slope associated with emotional stability was significant in all models except for OCB at time 2 (b = .019, SE = .026). None of the situational characteristic variables had significant effects, although this is not a surprising finding given the explanation mentioned previously. Again, the main effects of the situational characteristic variables convey their association with criterion-related validities aggregated for within each stem. Thus, a significant and positive effect for some situational characteristic predictor might be interpreted as suggesting that response option criterion-related validities tend to be larger in magnitude in stems high in the situational characteristic in question. As was noted earlier, however, between-stem variability in the response option criterion-related validities approximated zero and, hence, were constrained to zero in the models in Table 41. Because there is no between-stem variability in average response option correlations, there is no variability to account for by stem-level predictors.

Three of the interaction terms were significant: emotional stability and team task work for deviance at time 1 (b = -.051, SE = .023), emotional stability and interpersonal relations for deviance at time 2 (b = -.133, SE = .049), and emotional stability and morality/fairness for deviance at time 2 (b = .083, SE = .030). The interaction between emotional stability and social pressure and performance was also marginally significant for deviance at time 1(b = .129, SE = .067). Simple slopes for the interactions involved in deviance at time 1 are plotted in Figure 9.

With regard to the interaction between emotional stability and social pressure/performance for deviance at time 1, the slope of deviance on emotional stability was stronger and negative at lower levels of stem-level social pressure/performance. At low social pressure/performance, the estimated equation was $\hat{Y}$ = .007 – .096*X; at high social pressure/performance, the estimated equation was $\hat{Y}$ = .004 + .020*X, where $\hat{Y}$ corresponds to the predicted response option criterion-related validity in predicting deviance at time 1 and X is emotional stability trait expression. Thus, the effect of emotional stability trait expression on correlations with deviance at time 1 was attenuated among stems characterized as high in social pressure/performance, and became stronger and negative among stems not emphasizing social pressure/performance. Stem-level team task work had the opposite effect on the regression of deviance on emotional stability. At low levels of team task work, the estimated equation was $\hat{Y}$ = .006 + .002*X; at high levels of team task work, the estimated equation was $\hat{Y}$ = .006 – .077*stability, again where $\hat{Y}$ corresponds to the predicted response option-level criterion-related validity in predicting deviance at time 1 and X corresponds to emotional stability trait expression.

Simple slopes for the interactions involving deviance at time 2 are plotted in Figure 10. With regard to the significant interaction between emotional stability and interpersonal relations

in predicting response option-level criterion-related validities for deviance at time 1, the slope of for emotional stability was stronger and more negative among stems associated with higher scores on interpersonal relations. Among stems low in interpersonal relations, the estimated equation was $\hat{Y} = .004 + .019*X$; among stems high in interpersonal relations, the estimated equation was $\hat{Y} = .010 - .103*X$, where $\hat{Y}$ is the predicted response option-level criterion-related validity for deviance at time 2 and X is emotional stability trait expression. Thus, the effect of emotional stability trait expression on correlations with deviance at time 2 was attenuated among stems characterized as low in interpersonal relations, and became stronger and negative among stems high in interpersonal relations. Stem-level morality/fairness had the opposite effect on the regression of deviance on emotional stability. Among stems characterized as not emphasizing morality/fairness, the estimated equation was $\hat{Y} = .007 - .094*X$; among stems characterized as not emphasizing morality/fairness, the estimated equation was $\hat{Y} = .006 + .011*X$, where $\hat{Y}$ is the predicted response option-level criterion-related validity for deviance at time 2 and X is emotional stability trait expression.

In sum, limited support was provided for Hypothesis 3b. Of the few significant interactions observed in Table 42, all involved deviance; none of the interactions between trait expression scores for emotional stability and the situational characteristics was significant for the other outcomes modeled. As was suggested earlier in the stem-level models, the lack of significant effects is likely to be partially attributable to the relatively small number of stems available for the analyses involving some of the outcomes, which results in a reduced power for the cross-level interactions.

Summary of Results: Hypotheses 1-3b

Hypothesis 1 suggested that situational characteristics associated with item stems would account for between-stem variability in stem-score trait saturation, that is, correlations with other individual difference characteristics in the domains of personality, ability, and experience. Similarly, Hypothesis 2 predicted that situational characteristics would account for between-stem variability in stem-score criterion-related validities, that is, correlations with various criterion outcomes.

Models estimated to test Hypothesis 1 were mixed in terms of support provided for the proposition that stem-level situational characteristics would explain stem-score trait saturation. Few of the zero-order correlations between the eight situational characteristics under study and the trait saturation correlations reached significance, thus providing somewhat limited support for the relevance of individual situational characteristics. In particular, of the 64 correlations in Table 27, only three were significant at the conventional $p < .05$ level (ability and competition: $r = .301$; ability and interpersonal relations: $r = .241$; extraversion and morality & fairness: $r = -.242$). However, OLS models indicated that situational characteristics collectively accounted for approximately 3-23% of the between-stem variability in stem-score trait saturation associated with FFM personality characteristics, ability, and experience (Table 31). Significant effects were observed in the models predicting stem-score correlations with: ability (competition was positively related to stem-score correlations with ability), emotional stability (less familiar and more difficult situations were associated with stronger stem-score correlations with emotional stability), extraversion (competition was positively related to correlations with extraversion, whereas morality and fairness and social pressure and performance were associated with lower correlations with extraversion), openness (situations with greater demands associated with interpersonal relations and fewer demands associated with morality and fairness were associated

with stronger relationships with trait openness), and experience (number of business courses; stems that were more familiar and perceived as more easily managed by the average person had stronger relationships with experience).

As a whole, then, there is some support for the argument that ratings of situational features associated with SJT item stems can be used to predict stem-score trait saturation in that several of the relationships were significant and a non-negligible percentage of variability in correlations was explained by the models. However, trait saturation correlations with certain individual difference characteristics (e.g., agreeableness, conscientiousness, experience as measured via job tenure) were not explained by stem attributes. Further, additional research is needed to explore in greater detail relationships between specific situational attributes and specific individual difference characteristics.

At this juncture, perhaps the most definitive statement that can be made is that the measures of task demands and team task work were not useful in predicting stem-score correlations in that they failed to account for stem-score trait saturation correlations (as well as criterion-related validities). One possible argument is that there was insufficient between-stem variability in these characteristics, perhaps because of the way SJTs for educational and organizational contexts are developed (i.e., cues associated with tasks and working on tasks with others pervade these contexts). This argument is countered by the fact that many of the item stems, particularly for the College Board SJI, pertain specifically to social situations (e.g., events transpiring in a dormitory or at a party) where there is little to no emphasis on individual or team task cues. Furthermore, tests of between-stem variability (Table 16) provided inferential evidence support for the notion of between-stem variability in these characteristics; relatedly, the stem variance components for these characteristics (Table 14) were among the largest for all of

the situational characteristic scales. Therefore, the lack of predictive effects for task demand and team task work does not appear to arise because of insufficient stem-level variability in these characteristics.

Another possibility for the observed lack of effects for task demands and team task work is that task demands and team task work, as operationalized herein, may exist at too gross a level to be useful as predictors. In other words, task demands and cues associated with team task work might have to be differentiated at a greater level than was done herein. With these arguments in mind, there was some evidence that task demands and team task work moderated response option-level relationships. In particular, team task work interacted with emotional stability trait expression in predicting response option-level criterion-related validities for deviance (Table 42). Thus, although support for task demands and team task demands was not overly strong, it may be premature to conclude that these characteristics are unimportant.

With regard to stem-score correlations with criterion outcomes, situational characteristics collectively accounted for 10-32% of the between-stem variability in stem-score criterion-related validities (Table 32). Significant effects were observed in the models predicting stem-score correlations with: deviance at time 1 (morality and fairness was negatively related to correlations with deviance) and OCB at time 1 (more familiar and less difficult situations were associated with stronger stem-score correlations with OCB). As a whole, then, support was conflict for the argument that situational features associated with item stems can be used to predict stem-score criterion-related validities. On the one hand, few significant effects were observed in the regression models displayed in Table 32. However, the percentage of variance accounted for by the models was larger than that observed for the models for the individual difference correlations (Table 31). One potential explanation for the conflicting results is that sample size at the stem

level for the criterion models was simply too small to permit precise parameter estimation given the number of parameters in the models. This point is returned to in the Conclusion section with regard to study limitations.

Having examined the use of item stem characteristics for predicting stem-score correlations with other variables, the next step was to examine response option correlations with both individual difference characteristics and criterion outcomes prior to examining Hypotheses 3a and 3b. With regard to correlations with individual difference characteristics (i.e., response option-level trait saturation), models were first estimated for trait saturation involving the FFM personality traits given their analog with the FFM trait expression scores associated with the response options. In OLS models pooling over item stems, FFM trait expression accounted for approximately 4-20% of the between-option variability in correlations with like FFM characteristics. The strongest relationship observed was for conscientiousness, where a unit increase in conscientiousness trait expression was associated with an increase of .043 in the predicted response option-level correlation with conscientiousness. The relevance of a .043 increase is apparent when considering that a given SJT will be comprised of many response options. If a test were to be comprised of options that are relatively high in conscientiousness trait expression but are not overly redundant from the standpoint of the between-option correlations, then aggregating across these items would be expected to yield a composite score that might be fairly strongly related to trait conscientiousness.

The OLS models estimated for FFM trait expression predicting like FFM characteristics were extended by permitting the intercept and slope terms to vary across item stems. Evidence was found for clustering effects for the regression slopes; that is, the effect of FFM trait expression varied across the item stems examined herein. This variability in slopes represents an

interaction between the stems themselves, as discrete units, and the trait expression scores in predicting trait saturation. The percentage of variance in the response option-level correlations accounted for by the models permitting the slopes to vary (estimates ranging from .168-.404; Table 36) was substantially larger than for the models where the slopes were pooled over stems (estimates ranging from .039-.205; Table 33). Figure 4 depicts the dispersion of each of the five slopes as indicated by the relative width of the densities. Across stems, FFM trait expression slopes were almost uniformly positive for openness and conscientiousness; 98.9% of the slopes for openness and conscientiousness were greater than zero in magnitude. A similar pattern was observed for agreeableness, with 92.2% of the slopes being greater than zero in magnitude. A smaller percentage of the slopes were positive for emotional stability and extraversion, with around 78.9% of the slopes being positive for both. Thus, in spite of the overall positive relationship between FFM trait expression and response option-level FFM trait saturation correlations, variability in the slopes did result in some of the relationships approximating zero or being negative for some of the stems, particularly for emotional stability and extraversion.

In contrast to the regression slopes, no evidence was found for clustering effects associated with regression intercepts; adding a variance parameter to the intercept term did nothing to improve model fit (Table 35). Although such a finding might raise suspicion (e.g., that the lack of intercept variability was the result of some methodological decision such as the method of predictor centering), the lack of intercept variability was not simply an artifact of how the models were constructed or estimated. Table 43 shows the results of a means model[3]

[3] Estimates in Table 40 were generated from a model with stem set as a random factor and fixed-effects estimates generated for an indicator term corresponding to the variable with which the correlations were associated (e.g., each of the FFM, experience, ability, and the various criterion outcomes). The effect of the indicator variable was permitted to vary across stems, which yields

yielding, for each response option correlation, an overall mean estimate across all response options as well as the standard deviation of the correlations across stems. As Table 43 illustrates, there was no evidence of between-stem variability in the response option correlations with any of the variables under study. That is, there is no indication of clustering effects of the response option correlations by stem for either response option-level trait saturation or criterion-related validity; all variability in response option correlations was found across options within stems as opposed to between stems.

At first blush, a total lack of between-stem variability in response option correlations might seem peculiar, even if it's not of primary interest in and of itself. Rather, it would seem intuitive that response option correlations would vary across stems at least to some small degree, where if one were to average the correlations across response options within each stem, some stems would have higher means and others would have lower means. One potential explanation for a lack of between-stem variability in correlations pertains to how SJTs are designed. In particular, within a given stem, some response options might be expected to have relatively larger correlations with external variables whereas others have lower, or even negative, correlations with external variables. Consequently, when these correlations are averaged within each stem, the positive and negative signs on the correlations wash out.

This washing-out argument might be particularly relevant for SJTs that are scored using a forced-choice format, where respondents are prompted to choose or endorse only a subset of the response options for each stem (e.g., to select the best and worst). With a forced-choice format, zero-order correlations between endorsement of each response option and external variables

---

the *SD* estimates. The intercept of the means model was set to 0, providing estimates for the means for each correlation (if the intercept had been estimated, one of the levels of the indicator variable would be set as the reference level and the effects of the other levels would be deviations from the reference instead of as mean estimates).

might be constrained within a given stem because of an ipsative-like dependence that does not arise when each response option is rated independently of the others (e.g., using a Likert-type format). If such an explanation were true, then, one might expect to see zero or near-zero variability in response option correlations for the College Board and Managerial SJIs used in the present study, both of which rely on a forced-choice format, and non-zero variability in correlations for the Team Role Test, which utilizes a rating-scale format.

To explore this issue, the model used to generate the estimates in Table 43 was refit separately to the Team Role Test and then on both the College Board and Managerial SJIs. Between-stem estimates for response option-level correlations within the Team Role Test were small, but non zero, varying between .002 and .009 (by way of comparison, the residual *SD* estimate was .087, which is notably larger than the between-stem estimates). Conversely, all stem variance component estimates were zero among the College Board and Managerial SJI items. The non-zero estimates for the TRT suggest that response format may make somewhat of a difference in terms of mean/intercept variability, although perhaps not a particularly large one. It should be kept in mind that the TRT is only one measure and only contains ten stems; hence, caution should be exercised in generalizing to other rating-scale format SJTs given the small sample size and any potential design characteristics that are idiosyncratic to this measure.

In the full models containing situational characteristics and FFM trait expression predicting correlations with FFM personality characteristics, the slope of like FFM trait expression was always significant, whereas there was no evidence of main effects associated with the situational characteristics. Given the aforementioned lack of systematic between-stem variability in response option trait saturation correlations and criterion-related validities, the lack of significant situational characteristic effects is not surprising. Some evidence was found for

interactions between stem characteristics and response option trait expression in the prediction of trait saturation correlations for conscientiousness, emotional stability, and openness. In particular, cues associated with interpersonal relations were associated with more positive slopes in the model for conscientiousness trait saturation, whereas emotional demands and situations involving issues associated with social evaluation and judgment were associated with weaker slopes for conscientiousness trait saturation. For openness, more negative slopes were associated with situations that were more familiar, less difficult, and had cues demands associated with morality and fairness. Finally, slopes for emotional stability tended to be more negative in stems that emphasized issues associated with morality and fairness and that were more familiar and less difficult.

Concerning the models for ability and experience, FFM trait expression predicted correlations with ability and experience as measured by job tenure; for both, all five FFM characteristics had significant effects. There was little evidence of FFM trait expression effects on correlations involving experience as measured by number of business courses, with extraversion being the sole exception. Furthermore, there was little evidence of stem-level intercept or slope variability in the models predicting trait saturation correlations for ability and experience. In other words, although FFM trait expression characteristics associated with the response options tended to have significant relationships with trait saturation correlations for ability and experience, there was no evidence that parameters associated with these relationships were influenced by the item stems within which the response options were nested. Consequently, given the lack of intercept and slope variability in the models for ability and experience, situational characteristics were not modeled as predictors as stem-level predictors of response option-level trait saturation correlations with ability or experience..

Finally, models were estimated where criterion-related validities with various outcomes (OCB, GPA, overall performance as rated by BARS, deviance) were regressed on situational characteristics and FFM trait expression. The majority of the relationships between the individual trait expression scales and correlations with the outcomes were significant (Tables 37 and 38). Thus, the present study is the first to demonstrate that characteristics associated with situational judgment test content in terms of FFM trait expression characteristics can be used to predict the extent to which SJT response options predicted various outcomes. Given the large number of parameters that would have to be estimated for a model with eight situational characteristics and five FFM trait expression predictors, one FFM trait expression predictor, emotional stability, as this predictor demonstrated the greatest between-stem variability in regression slopes (Tables 38 and 39).

Significant interactions between situational characteristics and emotional stability trait expression were limited to deviance. The negative relationship between emotional stability trait expression and deviance at time one was stronger (more negative) for stems that emphasized cues associated with team task work. There was also some evidence that demands associated with social pressure (e.g., evaluation) and performance in social settings attenuated the negative relationship between emotional stability and deviance at time one. At time two, stem demands associated with interpersonal relations strengthened the relationship between emotional stability and correlations with deviance, whereas the effect of emotional stability appeared to be dampened in stems in which issues associated with morality and fairness were emphasized. Aside from deviance, there was little evidence that situational characteristics were related to emotional stability slopes and correlations with BARS or OCB.

CONCLUSION

The purpose of the present study was to examine whether information concerning characteristics of SJT stems and response options could be used to understand relationships between item-level response scores and correlations with external variables. More specifically, the present study examined the utility of item stem characteristics in terms of situational demands and features and response option characteristics in terms of FFM trait expression in explaining item-level correlations with other individual differences in the domains of personality, ability, and experience as well as various criterion outcomes (e.g., deviance, GPA). The remainder of the dissertation summarizes the study's strengths and weaknesses, discusses the study's implications with regard to situational judgment test design, and concludes with suggestions regarding future research.

Implications

The present study has implications for various areas of research and practice associated with the development of situational judgment tests and other types of situationally-based measurement procedures. From a practical perspective, researchers interested in applied measurement and assessment design have called for systematic, evidence-based approaches to test development. Such approaches are particularly relevant for measurement systems designed to assess multidimensional, complex characteristics and performances such as those that often motivate the use of situationally-based measurement procedures (e.g., Mislevy & Haertel, 2006; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002). Pressures that drive evidence-based, standardized approaches also stem from heightened demands for complex assessment

methodology in high-volume testing and increasing capabilities for simulating work environments (e.g., Mislevy, Steinberg, & Almond, 1999).

Mislevy and colleagues (1999; see also Mislevy, Almond, & Lukas, 2003) developed a conceptual model for evidence-based assessment design. Their approach encompasses various submodels pertinent to the development of complex assessment procedures, including considerations regarding how measurements yield operationalizations of the targeted performance or characteristic, considerations regarding the requisite technology and the environment within which assessment will occur, and so forth. With regard to test content, Mislevy and colleagues posited a task component that pertains to task design features. One of the key ideas underlying evidence-based design in terms of task design pertains to the identification of situational features that evoke behaviors of interest, which is of central importance for situationally-based measurement procedures such as SJTs.

In accord with this broad concept of evidence-based design, the present study was predicated on the idea of examining if and how information regarding the psychological features associated with SJT content might aid in understanding of properties of SJT scores and, ultimately, in systematic approaches toward SJT development. In particular, SJTs are apt for the measurement of behaviors (or, more specifically, judgments regarding behaviors) evoked in response to complex situations, as they are capable of depicting "ambient details" that reflect the types of situations to which one wants to generalize (Christian et al., 2010; p. 104). This argument, which extends to other simulation-based measurement procedures, motivates an understanding of those characteristics of SJT stimuli in terms of what contributes to the "ambient details" that Christian and colleague speak of. In conjunction with other recent research (e.g., Kell et al., 2010), the present study represents an initial step at systematically scaling thematic

content in SJTs based on theoretically-relevant characteristics with the aim of improving current knowledge that can be used as evidence in the design of SJTs and other situationally-based measurement procedures.

SJT development frequently begins with the collection of critical incidents that guide the generation of stem and response option content (Anderson & Wilson, 1997; McDaniel & Nguyen, 2001). This approach to test design produces content that is relevant to the job context on a surface level, which that has obvious potential benefits (e.g., evidence supporting content-valid test design, applicant reactions and perceptions associated with face validity). In general, however, this process does not include an explicit consideration of the psychological attributes (e.g., affordances, demands, inhibiting characteristics) associated with the incidents that are presented or how modifications to the incidents (e.g., altering aspects of the situation as presented in the item stem) might systematically affect these attributes. Rather than taking into account specific psychological features of the situations present in the job context, the focus is on sampling discrete, surface-level characteristics (e.g., interactions with customers) with the aim of including content that is relevant to some broad dimension (e.g., customer service). In the long term, a greater understanding of the relevant psychological features of the situations presented in SJTs and other situationally-based measurement procedures has the advantage of being able to more closely align current conceptualizations of content-valid design, based on the use of critical incident-like methods, with evidence-based design. This union has a number of benefits.

As is the case for assessment centers, work samples, situational interviews, and other situationally-based measurement procedures, SJTs designed for specific jobs contain situational content that is intended to be a sample of the content domain associated with behavior in the job in question. Thus, in contrast to measures of personality or attributes of various cognitive

domains (e.g., ability, aptitude, knowledge, skills), situations in SJTs are frequently not chosen so as to produce stimuli that will yield parallel indicators of a specific behavior or trait. Rather, the intent is to capture samples of respondent behavior that are representations of those in the job in question (e.g., Neidig & Neidig, 1984) with measures developed so as to maximize this correspondence between test performance and criterion performance (Asher & Sciarrino, 1974; Wernimont & Campbell, 1968). Hence, a test developer could use data concerning the situational characteristics of a given SJT to link to situational cues that exist in targeted jobs.

The measurement of situational cues in jobs could be accomplished via methods such as personality-oriented job analysis (e.g., Foster, Gaddis, & Hogan, 2012; Raymark, Schmit, & Guion, 1997; Tett & Burnett, 2003). Alternatively, one could also develop job analysis methods that are designed to be more sensitive to contextual and situational features relative to typical approaches that focus on worker requirements or job characteristics (see also Harman's [2012] discussion of context analysis as a work analysis methodology or framework). Such a methodology could be grounded in a model such as that used in the present study. The closer the approximation between the situational content in a simulation-based method such as an SJT and that found on the job, the more veridical the representations of likely performance on the job (Weekley & Jones, 1997) and, as a likely consequence, the more accurate the predictions that can be made from test scores (e.g., Wallace, 1966). Consequently, data regarding the overlap between the situational characteristics represented in test content and those found on the job could complement information derived from traditional critical incident-based approaches to drive validity arguments associated with content domain representation and relevance (e.g., Lennon, 1956; Messick, 1989) that are more specific, targeted, and perhaps procedurally replicable than what is capable via typical critical incident-based approaches alone.

In addition to supporting arguments associated with content representation and relevance, another design benefit of understanding situational characteristics associated with test content pertains to item development and selection. A common outcome of the critical incident generation process are groups of content-similar incidents within which one must decide what to retain for inclusion in the test (McDaniel & Nguyen, 2001). Because the items in each group are judged to be similar in terms of surface characteristics, the researcher might believe that the items within each group are largely interchangeable. However, evidence from expert panel reviews and statistical indices of response agreement suggests that this assumption is wrong; rather, very minor changes to SJT item content may influence how item content is interpreted (Clause et al., 1998), making choices among seemingly similar items more consequential than might be assumed when designing the test. A better understanding of how such content differences influence the psychological features associated with the incidents could be beneficial in developing pools of items that go beyond nominal, discrete, or surface similarity.

Knowledge of the psychological characteristics underlying SJT content could be leveraged in a number of potentially innovative ways. For instance, information concerning the psychological characteristics of situations or response options might reveal that the fidelity with which different features are represented depends strongly upon the medium of administration (e.g., video-based as opposed to a written format), whereas other features are represented equally well irrespective of format or medium. This possibility was speculated about in the Results section with regard to differences between the situational characteristic scales concerning the magnitude of the stem variance components estimates (i.e., large variance component estimates found for the Task Demands, Morality & Fairness, and Team Task Work scales and very small estimates found for the Individual Emotional and Familiarity & Difficulty scales). If medium is

meaningful in this manner, increases in fidelity associated with going from a written format to a video-based format or otherwise are conditional on the psychological characteristics associated with the content presented in the situations. If such increases are conditional on test content in this manner, there may be instances where the cost of developing video-based or higher-fidelity instruments would not be justified.

Another design application of knowledge regarding the situational characteristics of SJT content pertains to the use of tailored testing procedures and item branching. In particular, it is possible that information concerning psychological characteristics can be leveraged to inform the selection of items to be administered based on how the individual responds to earlier items. Doing so may thus help to gain a better understanding of how contingencies in a respondent's behavior are associated with specific situational features, or to probe into certain situation characteristics that seem problematic for the respondent. This idea was foreshadowed by Mischel (1973), who suggested that useful information about respondent characteristics could be obtained by systematically manipulating or altering features of the stimulus environment to observe how the individual's behavior changes in response. Such an approach seems particularly promising for SJTs and other situationally-based method procedures.

Finally, another operational application of knowledge regarding situational features pertains to parallel test form development with SJTs and other simulation-based methods. Clause et al. (1998) noted that unexplained multidimensionality inherent in measurement procedures such as SJTs complicates the development of alternate item banks that reflect the characteristics of the original target item bank. Therefore, they developed an item cloning procedure designed to replicate multidimensionality in the original form. Oswald and colleagues (2005), in turn, adapted the Gibson-Weiner procedure for developing parallel test forms to address similar

issues. The procedure used by Oswald and colleagues permits the test designer to leverage auxiliary or collateral information concerning item characteristics in order to inform parallel test form development. Oswald and colleagues applied this procedure by using item-GPA correlations to ensure that the alternate form would have criterion-related validity. In a similar vein, it may be possible to use information concerning situational characteristics inherent in the item content in an attempt to ensure that stimuli used in the alternate form were similar in standing on key dimensions to that of the original target test form.

In addition to the practical implications discussed above, the study also has theoretical implications for organizational research relevant to personality, particularly with regard to its applications to situationally-based measurement procedures such as SJTs and ACs. The present study examined specific features associated with situations and the traits expressed in response options to account for item-level variance in correlations with external variables (e.g., other individual-difference correlations, criterion-related validities). This approach is somewhat distinct from other recent research that has examined SJTs and ACs in light of personality or interactionist principles. First, the present study examined FFM trait expression as a response option-level attribute, whereas other researchers have examined trait-like characteristics at the level of situations. As an example, Tett and Guterman (2000) examined trait relevance, a situational characteristic defined and measured as the extent to which a situation would be expected to provoke the behavioral expression of a given trait in at least some people (p. 402). Trait activation theory, of which situation trait relevance is a central component, emphasizes the idea that situations vary in the types of trait-thematic cues that make up the situation, and that differences in the composition of cues influences the relevance and expression of traits across situations (Tett & Burnett, 2003; Tett & Guterman, 2002).

However, nothing in Tett and Guterman's (2000) actual operationalization of trait relevance addressed specific features of situations; rather, the researchers collected judgments of the relevance of the trait itself. There is nothing ostensibly wrong or incorrect with operationalizing trait relevance in this manner; so long as one acknowledges that actual cues, per se, were not measured, this approach to studying trait relevance even has certain advantages (e.g., it circumvents the field's lack of both a taxonomy for situational cues and theories regarding cue-trait linkages for specific classes of cues or traits). However, it could also be argued that this approach does not further the field's understanding of specific thematic features of SJT content, as judgments of trait relevance would be expected to be influenced by such cues (or would instead be conceptualized as trait content expressed in behavior instead of as a situational variable). Furthermore, researchers subsequent to Tett and Guterman (2000) have not always been careful in generalizing results from this study when discussing situational cues with regard to trait activation. For instance, Tett and Burnett (2003) interpreted the findings from Tett and Guterman as demonstrating that "…correlations between self-report trait measures and trait-relevant behavioral intentions are stronger in situations *providing appropriate cues for trait expression* … situations can vary reliably in the provision of *cues* for expressing target traits (i.e., trait relevance)" (italics added; p. 502). Given the emphasis placed on cues throughout Tett and Burnett's (2003) exposition of trait activation theory, it is surprising that such cues have not been directly examined.

Unlike Tett and Guterman and others that have examined trait-relevant characteristics at a level analogous to the situation (e.g., Haaland & Christiansen, 2002; Lievens et al., 2006), other researchers have examined trait expression as a characteristic of behaviors or response options. For instance, Kell and colleagues (2010) scaled SJT response options in terms of FFM trait

expression and linked trait expression ratings to raters' evaluations of the effectiveness of the response options in order to derive implicit trait policy (ITP) scores for each rater. Motowidlo and Beier (2010) examined the concept of differential attractiveness, a response option-level characteristic operationalized as the correlation between raters' evaluations regarding the effectiveness of an action and raters' standing on the personality characteristic in question. Like Kell and colleagues, Motowidlo and Beier (2010) also examined trait expression as a response option-level characteristic, linking it to response option differential attractiveness.

Both Kell et al. (2010) and Motowidlo and Beier (2010) examined ITPs, a construct that is central to Motowidlo and colleagues' ongoing theoretical and empirical research on knowledge constructs relevant to SJT scores. One way in which ITPs have been operationalized (Kell et al., 2010) is to regress each individuals' effectiveness ratings for a sample of response options onto FFM trait expression ratings for those response options. An individual's ITP for a given trait is operationalized as her or his weight from the regression equation associated with that trait. Results from the present study suggest that the effect of trait expression on correlations with other individual difference variables and criterion outcomes varied significantly across item stems, as evidenced by the significant improvement in model fit when slopes were permitted to vary across stems. Although the present study did not examine ITPs, the finding of stem-level dependencies does raise the possibility that variance in ITPs may also exist at the level of individual item stems and that approaches to estimating ITPs such as that used by Kell and colleagues ignores this variability.

Such stem-level variability in ITPs would imply that that how an individual weighs a given personality characteristic when evaluating the effectiveness of a series of response options depends on some aspect or aspects of the situation. The question then becomes one of explaining

patterns in the ITPs. One approach might be to proceed by categorizing ITPs via some method at the rater level of analysis (e.g., factor analysis) to ascertain patterns with regard to the content of stems that are clustered together. That an individual might have multiple ITPs for a set of situations is not at all at odds with the theory underlying ITPs. ITPs are conceived of as skills, habits, preferences, and the like that people adopt over time as they experience the world (Motowidlo & Beier, 2010). It is likely individuals vary in terms of their level or types of experience with specific situations, and this variability might itself cause an individual to hold multiple ITPs. Furthermore, within a class of situations that comprise a given setting (e.g., in a specific workplace), the situations themselves are bound to vary in terms of demands, affordances, and the like, and this variability is also likely to cause within-person heterogeneity in ITPs for a given trait.

Strengths, Limitations, and Suggestions for Future Research

Any study has strengths and limitations which must be considered when interpreting findings and attempting to generalize beyond the particulars at hand. Concerning strengths, the study examined three different SJTs designed for use in different contexts (i.e., admissions in post-secondary education, managerial selection, and team selection contexts). Being able to examine multiple tests increases heterogeneity in test content and, relatedly, reduces dependency of the results on idiosyncratic features of any single instrument. A second strength of the study pertains to the estimation and handling of stem-level dependencies in the response option-level models. Incorporation of stem-level effects in response option-level models accounts for systematic differences between stems in response option effects that would be ignored in an analysis that pools completely over stems. Furthermore, the design permitted for the examination

of how stem attributes in terms of situational features were related to response option-level effects in terms of trait expression on response option level correlations. Relatedly, the design also permitted for examination of two common types of SJT scoring protocol, namely stem scoring and response option-level scoring.

The pilot study of the situational characteristics and trait expression scales permitted the opportunity to estimate the number of raters that would be required for aggregation across raters given the levels of agreement and reliability specific to the present context (i.e., evaluations of SJT content). In most cases, it was found that 10-15 raters were sufficient to obtain adequate levels of agreement and reliability to justify aggregation across raters. A final strength of the present study pertained to the development of the Situational Characteristics Inventory (SCI). As there is no widely accepted, comprehensive taxonomy or model of situational features content relevant to organizational settings, the SCI is a tool for measuring situational content of assessment methods such as SJTs, ACs, work samples, and situational interviews. The scale might also be adaptable for settings where one is charged with assessing the presence of specific situational demands in a work analysis context.

In spite of the aforementioned strengths, a number of limitations and weaknesses are also apparent; in some cases, these relate to the strengths listed above. For instance, in spite of being able to examine multiple SJTs, some of the analyses conducted in the present study were likely to be underpowered given the number of the stems available (90) and the magnitude of the effects that were observed. Given the relative lack of prior research on which to gauge the likely magnitude of effects that would be observed prior to the study, there was little basis for accurately estimating the number of required stems that would be needed to address concerns regarding precision and uncertainty in parameter estimation in an a priori fashion. This problem

was exacerbated by the fact that only a subset of the 90 stems had data for some of the outcomes that were studied (e.g., correlations with experience; criterion-related validities). Going forward, researchers conducting similar studies should have more than 90 stems available if the item stem is a relevant unit of analysis. In the author's opinion, this will be a somewhat of a logistical challenge, as SJTs are often relatively short in length, in terms of number of item stems, given the time demands associated with SJT administration (e.g., time required for reading or viewing the critical incidents and associated response options).

A methodological limitation of the present study pertains to the relatively low ICC estimates observed for the Familiarity & Difficulty and Individual-Emotional situational characteristic scales. Although the estimate for the Individual-Emotional scale was somewhat low (.585), it is possible that adjustments to item content, the appending of additional items, or the use of additional raters could address the low observed ICC estimate. More concerning was the estimate for Familiarity & Difficulty (.394). Although the magnitude of the stem-level variance was larger for Familiarity & Difficulty than for Individual-Emotional (.060 versus .046), the rater and residual variance components for Familiarity & Difficulty were 2.5-3 times larger than that for Individual-Emotional (.439 versus .142 for the rater estimate; .501 versus .195 for the residual estimate). Although the Familiarity & Difficulty items were written to reference other people as opposed to the individual respondent (e.g., "The average person has dealt with situations similar to this in the past"), it is possible that ratings on these items are more indicative of some characteristic or phenomenon idiosyncratic to the individual rater than are items for the other situational characteristic scales. Such individual-specific idiosyncrasy does not negate familiarity or difficulty as also being stem-level attributes; there is nothing inconsistent in saying that some situations are perceived to be, on average, more familiar or

difficult than others even if individuals vary widely on how familiar or difficult they perceive the situations. The presence of individual-specific idiosyncrasy does, however, place greater demands in terms of the number of raters that are required to generate stable estimates of stem-level standing on this attribute. Another possibility unrelated to the individual-specific idiosyncrasy argument is that the stems examined herein may simply be subject to selection effects stemming from prior screening during the test development process, such that stems that are extremely low or high in familiarity or difficulty were not included in the tests.

Another potential methodological weakness pertains to the use of undergraduate students to provide the ratings of situational characteristics and FFM trait expression. One might argue that the study should have relied on ratings generated from SMEs such as those with backgrounds in industrial/organizational psychology or related disciplines or individuals with expert knowledge of the contexts in question. The demands inherent in the rating process in terms of the number of ratings that had to be made by each rater and the time required to carry out the rating task required a relatively large sample of raters, on the order of several hundred for the situational and FFM trait expression characteristics, which precluded the use of a small SME sample to provide the ratings. Relatedly, the lack of a formal training process prior to administration of the rating task might also be criticized, although raters were provided with detailed instructions that included an example of how to carry out the rating procedure. It remains for future research to examine the potential gains that might be made by including either actual SMEs for the rating process or conduct various types of rater training that might increase the quality of the ratings provided by either SMEs or non-SMEs.

Limitations could be noted with regard to the actual characteristics examined in terms of situational characteristics and response option trait expression. For instance, ratings could have

been collected concerning the expression of other characteristics or attributes aside from FFM personality features. Examples of such attributes might include general or specific abilities, knowledge, or experience. The adjectives used for the FFM trait characteristics herein (see Table 8) are, arguably, more concretely defined and less open to interpretation than would seem adjectives such as general ability or intelligence or knowledge, which are more abstract and, again arguably, more difficult to reference concisely with a one- to two-word description. Concerning the situational characteristics, an attempt was made to sample a broad array of features from the personality and organizational literature, neither of which has a comprehensive, organized framework for situational content. Even a framework such as the O*NET work context descriptors is broader than desired as it often pertains to stable features of the work environment or occupation as opposed to specific features of situations.

Finally, the item-level response data from which the correlations with individual difference characteristics and criterion measure were estimated came from research samples wherein the measures were administered for non-operational purposes. There is some evidence that correlations between SJTs and other variables may differ in incumbent versus applicant samples. For instance, MacKenzie and colleagues (2010) found that correlations between SJT scores (keyed using the Motowidlo et al. [1990] method of producing stem scores) and measures of cognitive ability were stronger in magnitude with incumbent versus applicant samples. Some differences were also found in the Mackenzie et al. study with regard to correlations between SJT scores and facet-level scores on personality measures (e.g., concern, cooperative, energy, initiative), although patterns of such differences were generally not consistent. In either case, to the extent that differences do exist with regard to correlations between stem scores or response option scores and external variables across contexts or settings, there is the potential that the

correlations used as outcomes in the present study were biased if one's intent is to generalize the results to operational, applicant samples. With that possibility in mind, unless there is heterogeneity in the effect of contextual or sample characteristics on correlations with external variables across items (i.e., not all items are similarly affected), then such a bias might not present a difficulty in generalizing the results found herein in terms of the relationships that were found. This, however, is an empirical question that remains open.

Considering the strengths and weaknesses noted above in conjunction with the findings of the study, a number of suggestions can be made for future research. First, a number of other measurement procedures used in personnel practice also rely on the sampling and replication of situational features from the context in question, including work samples, situational interviews, and assessment centers. Some researchers have explored the issue of whether psychological features in terms of situational and behavioral characteristics can be used to shed light on known measurement properties of some of these methods (e.g., Highhouse & Harris, 1993; Haaland & Christiansen, 2002; Kell et al., 2010; Lievens et al., 2006). The situational characteristics examined in the present study might similarly be useful for describing measurement procedures other than SJTs in order to inform design decisions or aid in the understanding of measurement properties. If so, they would provide a common framework used for examining content across specific types of situationally-based measurement procedures, which is advantageous for various reasons (e.g., aggregating across studies).

Second, the present study sought to examine, at a broad level, situational characteristics as a class of attributes relevant to SJTs. There may be value in focusing on some subset of the specific attributes examined herein at a finer-grained level of detail. That is, instead of examining the utility of a broad class of characteristics, one might choose instead to focus on a smaller

number of attributes on the basis of some criterion (e.g., relevance, theoretical interest, etc.) and measure them in a more precise manner than could be achieved in the present study. Third, the present study focused on written SJTs. Given that video and multimedia SJTs are also frequently applied in practice, the study of situational characteristics and FFM trait expression should also be relevant for tests administered using these media. In some respects, situational characteristics and trait expression may be more easily conveyed in a video-based format than with a written test, given the added richness of video as a medium in terms of the speed with which information can be conveyed and processed as well as the ability to include subtler nuances that cannot be easily or clearly conveyed in a written format. For both written and video SJTs, it would also be of interest to see how and whether content of the incidents can be systematically manipulated in some way to alter perceived situational characteristics or trait expression. Indeed, if one of the end goals of understanding psychological features of SJT content is to aid design decisions, then modification of content to influence these characteristics in a relatively predictable, systematic manner must be something that can actually be done. If it is, it has applications in SJT content development and potentially other areas (e.g., parallel test form development, computer adaptive test design).

Third, one observation that came about in examining the individual stem slopes in the regressions of FFM trait saturation on FFM trait expression was the existence of some stems with negative slopes (see Table 34 and Figure 3). That there exist different patterns across stems in terms of within-stem variability raises an important point about SJT item construction. Based on the author's experience with SJTs (including the present study), the construction of a set of stems and response options yields a series of response options that generally vary from one another in one of two ways. First, response options within a stem are sometimes written such that

each option expresses a different trait or characteristic. Thus, across a series of response options within a stem, one response might express a behavior relevant to honesty, another relevant to surgency, another relevant to agreeableness, and so forth. Such a series of response options would be viewed as heterogeneous with respect to the expressed traits. Second, response options are sometimes written to express one trait or dimension but at varying levels. Thus, one option might express a behavior that is low in accountability, another option that is moderate in accountability, and another that is high in accountability. Such a series of options would be viewed as homogeneous with respect to the expressed traits.

In reality, this heterogeneous/homogeneous distinction is not so likely a dichotomy as it is a continuum. In either case, it does have implications for some of the findings obtained from the present study. For instance, if all three of the tests examined herein tended to be written with response options that were heterogeneous in trait content within a stem rather than emphasizing differences in level, it might have implications for findings regarding response option trait expression variance components and ICCs. If one attempted to generalize these estimates to an SJT where response options are homogeneous with regard to trait expression within a stem, the estimates may not be appropriate. Although not taken into account in this study, additional research into this heterogeneous/homogeneous design consideration would be useful. At the test level, it could be coded as a study design characteristic and examined as a moderator in a meta-analytic study of SJT criterion-related validity or convergence with other characteristics. Item level approaches, such as that taken herein, could also be undertaken.

Fourth, one potential long-run implication from the present study is that SJT stems or response options might be constructed using systematic design principles associated with thematic content taken into consideration, as has been mentioned previously. The present study

examined stems and response options as separate or distinct units; that is, comparisons were made across stems or response options (a between-stem or between-option analysis). A logical next step is to examine the effects of systematic altering SJT content in an attempt to manipulate situational or behavioral characteristics in a predictable manner. In other words, if one has a stem that has some standing in its emphasis on some situational attribute (e.g., a moderate standing with regard to interpersonal struggles pertaining to power or resources), how can the stem be modified or altered in a way that might increase or decrease the stem's current standing (e.g., to yield a stem that strongly emphasizes resource struggles). Such a study necessitates a repeated-measures design where the same stem might be manipulated or altered and administered to examine the effects of the manipulation. As noted by Pervin (1978), the perception of situations may have gestalt characteristic. If so, it may be difficult to alter some specific cue or feature without also inadvertently changing others (e.g., increase the level of interpersonal struggles regarding resources might also effect the emotional demands associated with the situation).

Finally, situational and behavioral characteristics were examined herein with the goal of understanding item-level correlations with other individual difference characteristics and criterion outcomes. There are other phenomena that could be addressed via investigation of situational or behavioral characteristics of SJTs. Examples include response distortion, subgroup differences and related statistical concerns affecting test fairness (e.g., measurement invariance, predictive invariance), or the applicability of findings regarding situational and behavioral characteristics to develop new item keying methods.

In conclusion, the present study examined whether psychological features associated with situational judgment test content (namely, situational demands and attributes associated with item stems and personality trait expression associated with response options) could be used to

explain and understand heterogeneity in item-level correlations with other individual difference characteristics and relevant criterion outcomes. The present research was motivated by the belief that a greater understanding of these features of SJT content can be used to drive informed decisions regarding test design, development, and delivery. Although the models accounted for a sizable proportion of variance in item-level correlations, there were not enough instances of systematic patterns of relationships to state anything definitive at this stage regarding the potential usefulness of knowledge regarding psychological characteristics. Additional research with a larger number of items or focusing on a subset of the characteristics examined herein might provide fruitful for progressing our understanding of the psychological features of SJTs and other situationally-based measurement procedures.

APPENDICES

Table 1.

Empirically- and Theoretically-Derived Psychological Characteristics of Situations.

Study	Characteristics
Battisch & Thompson (1980)	***Emotional involvement***, characterized by intimacy ***Group versus individual activity***, characterized by social activities with friends versus being alone or working individually in a group setting ***Social isolation***, associated with knowing versus not knowing how to behave, involvement, assertiveness of behavior, and affective reactions associated with feelings of security versus insecurity, self-conscious versus at ease, and relaxed versus tense
Block and Block (1981)	***Structure***: situations characterized by well-defined roles, tasks, and goals ***Convergency****:* cognitively-oriented situations where the goal, task, or problem permits only one single correct answer ***Divergency****:* situations where the goal, task, or problem permits an open-ended number of alternate solutions ***Evaluation****:* situations where the accuracy, appropriateness, or desirability of one's behavior is understood to be evaluated by another party in a position of status ***Feedback****:* situations where information concerning the effectiveness, appropriateness, or desirability of one's behavior is provided by another party or is readily available through the observation of one's own efforts ***Constraint****:* situations where the goal, problem solution, or social interaction is constrained by the presence of a physical or psychological barrier

Table 1 (cont'd)	***Impedance**:* situations requiring a high degree of exertion that is affective, cognitive, or physical in nature ***Malleability**:* situations which permit change or restructuring by the actor
	***Galvanization**:* situations that are attractive to the actor or that have incentive value where the average person in the situation will be motivated and engaged
	***Familiarity**:* situations where the cultural, physical, or interpersonal context together with the task and social demands within the situation are known to the average person
	***Differentiation**:* situations that are highly articulated with number of "discriminanda" or regions
Eckes (1995)	***Nonintimate*** ***Emotionally uninvolving*** ***Informal*** ***Relaxed*** ***Social encounters*** ***Familiar social*** ***Frightening*** ***Emotionally involving*** ***Competitive* or *task-oriented***
Ekehammar and Magnusson (1973)	***Self-esteem* or *ego-threatening*** ***Positivity*** ***Social*** ***Activity*** ***Physical pain* or *threat***

Table 1 (cont'd)

Edwards and Templeton (2005)	***Valence:*** Extent to which the situation is positive or negative, depending on the extent to which the situation in question resulted in favorable or unfavorable outcomes for the respondent ***Task oriented****:* associated with goal achievement, productivity, or goal pursuit ***Effort*** (negotiation or routineness): extent to which respondents generally know how to behave in the situation or have the required skills to manage the situation's constraints
Endler, Hunt, and Rosenstein (1962)	***Threat to personal standing*** ***Personal danger*** ***Ambiguous situations***
Fleeson (2007)	***Anonymity****:* the number of others present in the situation, how familiar the respondent is with those individuals, and how much the individual like the others present in the situation ***Task orientation****:* perceived obligation, being evaluated, imposition, how close deadlines are, and the amount of interest in the situation ***Friendliness****:* how friendly others are in the situation, how much the participant interacted with the others in the situation, and the status of the others in the situation ***Status of others***
Forgas (1976)	***Perceived intimacy, involvement, and friendliness*** ***Self-confidence or perceived competence*** in light of the routineness of the situation

Table 1 (cont'd)	
	Evaluation: Extent to which the situation is evaluated positively or negatively; associated with situational constraint
Forgas (1983)	***Self-confidence****:* how self-confident the individual feels, knowing how to behave or not knowing how to behave, and feeling relaxed versus tense in the situation ***Evaluation****:* the extent to which respondents have a positive or negative affect toward the situation ***Seriousness****:* associated with perceptions of the situation being superficial versus intense, not serious versus serious, and simple versus complex ***Involvement****:* associated with the extent to which respondents perceived the situation as being involved versus uninvolved and intimate versus nonintimate
Frederiksen, Jensen, and Beaton (1972; see also Frederiksen, 1972)	***Evaluation of procedures*** ***Routine solutions*** ***Solution of interorganizational problems*** ***Solution of personnel problems*** ***Change in policy*** ***Conflicting demands on staff time***
Magnusson (1971)	***Positive* and *rewarding****:* e.g., receiving praise, performing well in spite of difficult circumstances ***Negative****:* e.g., receiving criticism, performing poorly, facing failure ***Passiveness****:* situations that required being idle, resting, or waiting ***Social****:* spending time with others and working with others ***Activity***

Table 1 (cont'd)

Pervin (1976)	***Friendly/unfriendly*** ***Tense/calm*** ***Interesting/dull*** ***Constrained/free***
Price and Bouffard (1974)	***Situational constraint****:* the extent to which the situation was "loaded" in terms of potential embarrassment for the respondent, whether the situation would require self-monitoring on the part of the respondent, whether *one's behavior in the situation would be affected by the approval or disapproval of others, and whether the situation demands certain behaviors over others*
Reis (2008)	***Dependence*** of the respondent's outcomes on the actions of others in the situation ***Distribution of power*** among actors in the situation ***Correspondence or conflict*** among actors Need to ***coordinate behavior*** among actors Whether the situation entails interaction among parties that will last for a ***short or long period of time*** ***Uncertainty***
Schutte and colleagues (1985)	***Prototypicality*** ***Behavioral constraint***
Sherman and colleagues (2010)	***Social*** ***School work in class with others*** ***School work at home or alone*** ***Recreating*** ***Getting ready for something*** ***Work***

Table 1 (cont'd)

	Unpleasant situations
Van Heck (1984; 1989; Van Heck, Perugini, Caprara, & Froger, 1994)	***Interpersonal conflict*** ***Joint working, exchange of thoughts, ideas, and knowledge*** ***Intimacy and interpersonal relations*** ***Recreation*** ***Traveling*** ***Rituals*** ***Sport*** ***Excesses*** ***Serving*** ***Trading***
Wish, Deutsch, & Kaplan (1976)	***Cooperative and friendly versus competitive and hostile****:* harmony versus clashing between parties, cooperative versus competitive, friendly versus hostile, and compatible versus incompatibility in goals. ***Equal versus unequal distribution of power****:* equality in power among parties in the situation as well as similarity with regard to roles and behavior ***Intense versus superficial****:* relations being characterized by activity versus inactivity and intense versus superficial interaction and feelings between parties ***Socioemotional versus task oriented and formal****:* emotional closeness between parties, sincerity versus insincerity, flexibility versus rigidity, and ease of ending contact between parties
Yang, Read, & Miller (2006)	***Valence****:* positive connotations of situations, interpreted as being associated with success in goal pursuit

Table 1 (cont'd)

	Goal achievement
	Failure
	Strong social bonds
	Being overwhelmed
	Lack of vision
	Being in danger
	Being morally or ethically challenged
	***Starting out* (goal initiation)**
	Turning bad to good
	At a standstill
	Being threatened
	Having no resolution
	Having necessary skills
	Achieving with ease
	Being socially inappropriate
	Deception
	Separation
	Having conflicting interests
	Enduring humiliation or embarrassment
	Making up for previously bad behavior
Wright and Mischel (1987)	***Cognitive demands****:* Rational thinking, short-term memory, knowledge, and intelligence ***Self-regulatory demands****:* Stress tolerance (tolerating frustration), ability to focus in the face of distraction, and delaying gratification ***Social demands****:* Verbal communication, dealing with conflict among peers, speaking in front of others, introspecting, and trying new situations

Table 1 (cont'd)	
	Physical demands*:* Encompassing physical strength, gross motor control, physical speed or quickness, and toughness
Mischel (1973)	***Situational strength***: Strong situations are those wherein uniform response patterns are observed among individuals, leaving no room for individual differences to influence behavior; weak situations are those situations where individual differences to manifest themselves.

Table 2.

Example Dataset Illustrating Between-Stem Analyses for Psychological Characteristics of Situations (n = 100 stems).

stem	*stem_f1*	*stem_f2*	*stem_f3*	*stem_f4*	*stem_f5*	*r_cons*	*r_agr*	*r_ext*	*r_open*	*r_neu*	*r_gma*	*r_perf*	*r_sat*	*r_abs*
1	3.15	6.34	1.00	5.90	2.17	.38	.16	.11	.21	.39	.23	.19	.13	.23
2	2.02	6.59	6.62	6.63	1.94	.35	.31	.37	.39	.21	.42	.24	.19	.31
3	5.11	6.39	3.90	3.09	1.76	.21	.44	.32	.38	.30	.14	.39	.24	.40
4	5.41	6.10	4.51	4.78	4.77	.31	.12	.44	.10	.38	.44	.32	.28	.24
5	2.57	4.88	5.49	3.05	1.18	.28	.20	.42	.14	.37	.35	.36	.27	.23
6	6.04	1.75	1.71	6.10	4.00	.29	.11	.11	.11	.19	.18	.32	.23	.32
7	5.72	1.97	1.52	6.59	6.34	.29	.40	.31	.27	.25	.10	.26	.34	.23
8	6.52	1.83	1.41	2.57	2.83	.21	.15	.36	.26	.24	.40	.29	.31	.10
9	2.85	5.20	1.52	5.86	4.71	.34	.28	.31	.30	.34	.35	.27	.42	.12
.														
.														
.														
100	3.72	6.86	1.00	2.44	4.11	.20	.13	.18	.21	.30	.45	.16	.29	.16

Note. stem_f1 through stem_f5 reflect situational features of stems, r_cons through r_neu reflect zero-order correlations between item stem scores and personality trait scores (cons = Conscientiousness, agr = Agreeableness, ext = Extraversion, open = Openness, neu = Neurotocism), r_gma reflects the zero-order correlation between item stem scores and cognitive ability scores, and r_perf through r_absent reflection zero-order correlations between item stems and relevant criteria (performance, satisfaction, and absenteeism).

Table 3.

Example Team Role Test (TRT) Item (Adapted from Mumford et al., 2008).

You are a member of a sales team at a local bookstore, where recent sales have been decreasing substantially due to a shrinking number of customers. You are in a team meeting discussing solutions to the declining sales problem. The discussion becomes a bit heated when the oldest team member suggests that the sales numbers for the new sales reps are quite low. One of the younger reps quickly counters that every time he asks for help with a customer, the older rep takes credit for the sale. The other new sales rep simply looks at the floor and says nothing. Please rate the effectiveness of each of the following responses.

	Very Ineffective	Somewhat Ineffective	Neutral	Somewhat Effective	Very Effective
Get the quiet new sales rep involved by asking if she has noticed that the older sales rep has taken some of her sales as well (role inconsistent).	1	2	3	4	5
Remind the two sales reps that personal attacks are not appropriate and that the team should focus on the future solutions.	1	2	3	4	5
Support the new team members by taking their side to make sure they are not used as "scapegoats" for the team's problems (role inconsistent).	1	2	3	4	5
Remind the team that making critical remarks about specific people makes people defensive and will prevent the members from accomplishing anything as a team.	1	2	3	4	5

Table 4.

Example Dataset Illustrating Between-Response Option Analyses for Psychological Characteristics of Situations (n = 100 stems).

stem	*ro*	*stem_f1*	*stem_f2*	*stem_f3*	. .	*ro_agr*	*ro_con*	*ro_ext*	. .	*ro_r_agr*	*ro_r_con*	*ro_r_ext*	. .
1	a	3.15	6.34	1.00	. .	1.96	6.73	4.65	. .	.45	.42	.61	. .
1	b	3.15	6.34	1.00	. .	3.09	5.90	1.69	. .	.45	.47	.43	. .
1	c	3.15	6.34	1.00	. .	4.70	2.98	2.17	. .	.36	.23	.36	. .
1	d	3.15	6.34	1.00	. .	2.02	4.76	2.26	. .	.32	.27	.35	. .
1	e	3.15	6.34	1.00	. .	1.65	4.42	5.27	. .	.34	.59	.58	. .
2	a	2.02	6.59	6.62	. .	4.55	6.78	4.81	. .	.29	.43	.19	. .
2	b	2.02	6.59	6.62	. .	5.14	5.73	2.17	. .	.52	.43	.23	. .
2	c	2.02	6.59	6.62	. .	5.81	6.03	5.67	. .	.10	.11	.11	. .
2	d	2.02	6.59	6.62	. .	5.24	4.52	6.04	. .	.08	.27	.70	. .
2	e	2.02	6.59	6.62	. .	4.68	4.88	1.93	. .	.30	.24	.12	. .
3	a	5.11	6.39	3.90	. .	4.41	6.62	6.46	. .	.39	.19	.31	. .
.	.	.	.	.	. .	.	.	.	. .	.	.	.	. .
.	.	.	.	.	. .	.	.	.	. .	.	.	.	. .
100	a	3.72	6.86	1.00	. .	5.24	5.89	1.82	. .	.06	.03	.47	. .
100	b	3.72	6.86	1.00	. .	3.71	3.25	4.83	. .	.23	.50	.37	. .
100	c	3.72	6.86	1.00	. .	1.83	3.56	2.63	. .	.61	.24	.66	. .
100	d	3.72	6.86	1.00	. .	6.29	2.66	6.60	. .	.32	.60	.02	. .
100	e	3.72	6.86	1.00	. .	1.89	5.02	6.43	. .	.32	.24	.33	. .

Note. Periods in table mark abbreviated rows or columns. *ro* = response option, *ro_agr* = response option rating for trait Agreeableness, *ro_con* = response option rating for trait Conscientiousness, *ro_ext* = response option rating for trait Extraversion, *ro_r_agr* = zero-order correlation between response option score and trait Agreeableness, *ro_r_con* = zero-order correlation between response option score and trait Conscientiousness, *ro_r_ext* = zero-order correlation between response option score and trait Extraversion.

Table 5.

Illustration of Nested Design Based on Hypothetical Example of 25 Situations Rated by 25 Raters.

	SJT Item Stems																								
Rater	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
1	x	x	x	x	x																				
2	x	x	x	x	x																				
3	x	x	x	x	x																				
4	x	x	x	x	x																				
5	x	x	x	x	x																				
6						x	x	x	x	x															
7						x	x	x	x	x															
8						x	x	x	x	x															
9						x	x	x	x	x															
10						x	x	x	x	x															
11											x	x	x	x	x										
12											x	x	x	x	x										
13											x	x	x	x	x										
14											x	x	x	x	x										
15											x	x	x	x	x										
16																x	x	x	x	x					
17																x	x	x	x	x					
18																x	x	x	x	x					
19																x	x	x	x	x					
20																x	x	x	x	x					
21																					x	x	x	x	x
22																					x	x	x	x	x
23																					x	x	x	x	x
24																					x	x	x	x	x
25																					x	x	x	x	x

Table 6.

Domains of Situational Characteristics.

O*NET Work Context
- *Consequence of Error*
- *Importance of Being Exact or Accurate*
- *Impact of Decisions / Work on Co-Workers or Company Results*
- *Structured versus Unstructured Work*
- *Time Pressure*
- *Level of Competition*
- *Work with Work Group or Team*
- *Coordinate or Lead Others*
- *Contact with Others*
- *Deal with External Customers*
- *Deal with Aggressive People*
- *Deal with Unpleasant or Angry People*
- *Conflict Situations*
- *Responsibility for Outcomes and Results*
- *Decision Making*

Work Design Characteristics (Morgeson & Humphrey, 2006)
- *Motivational – Autonomy*
- *Motivational – Task Identity*
- *Knowledge – Job Complexity*
- *Knowledge – Information Processing*
- *Knowledge – Problem Solving*
- *Knowledge – Specialization*
- *Social – Social Support (Emotional and Informational)*
- *Social – Interdependence*
- *Social – Interaction outside the Organization*
- *Social – Feedback*

Situational Demands (Wright & Mischel, 1987)
- *Cognitive Demands*
- *Self-Regulatory Demands*
- *Social Demands*

Customized / Adapted
- *Socioemotional Threat*
- *Emergency Situations*
- *Ethical*
- *Status / Power*
- *Emotional or Behavioral Constraint and Expression*
- *Stress*
- *Emotional Reactions*
- *Fairness*
- *Uncertainty*
- *Diversity*
- *Convincing / Persuasion*
- *Resources*
- *Being Evaluated and/or Recognized*
- *Others*
  - *Trust / Self-Interest*
  - *Familiarity & Challenging*
  - *Valence (Positive, Negative)*
  - *Involvement & Activity*
  - *Comfort*

Table 7.

Situational Characteristic Inventory (SCI) Item Content.

Instructions. You will be presented with descriptions of five situations. After each situation, you will be presented with a series of statements that describe everyday situations in general. Some of the statements will appear to be very relevant to describing the situation you've read, whereas other statements may appear to be not very relevant at all.

First, read the situation. Note that each situation concludes by asking you what you might do if you were in the situation. Your task, however, is ***not*** to indicate what you would do. Rather, you are being asked to rate each descriptive statement in terms of the extent to which it is **relevant for describing** the situation in question, from *Very Irrelevant to the Situation* to *Very Relevant to the Situation*. Some statements will pertain to ***other people*** in the situation (e.g., what *others* are doing, or *someone else* is doing in the situation). Other questions ask you to describe what ***you as the actor or participant*** (participant is abbreviated as *P* in some of the statements) are doing, would do, or so forth.

	Very Irrelevant To Situation	Somewhat Irrelevant to Situation	Neither Relevant Nor Irrelevant to Situation	Somewhat Relevant to Situation	Very Relevant to Situation
1. Situation requires intellectual capacity, verbal fluency, or rational thinking in order to resolve an issue or problem (examples: technical reports being reviewed, intellectual conversation among actors, a complex problem to solve).	1	2	3	4	5
2. Situation requires that one manage her or his emotions in order to resolve an issue or problem (examples: tolerate frustration or remain calm).	1	2	3	4	5

Table 7 (cont'd)

3. Situation demands that one monitor progress in achieving important goals or objectives in order to resolve an issue or problem.	1	2	3	4	5
4. Situation entails interacting with others and requires social skills in order to resolve an issue or problem (example: making a good impression, communicating tactfully and with consideration).	1	2	3	4	5
5. Situation is one in which the consequences associated with an error or mistake are high.	1	2	3	4	5
6. Situation demands precision, accuracy, or attention to minor details.	1	2	3	4	5
7. Situation contains formally imposed structure (examples: rules, deadlines, close supervision, clear roles, routine tasks).	1	2	3	4	5
8. Situation contains a strict deadline or working quickly under time pressure.	1	2	3	4	5
9. Situation involves competition between persons or groups.	1	2	3	4	5
10. Situation involves coordinating others' work or leading others to accomplish activities.	1	2	3	4	5
11. Situation involves a person or group acting in an aggressive or hostile manner.	1	2	3	4	5

Table 7 (cont'd)

12. Situation requires dealing with unpleasant, angry, or discourteous people.	1	2	3	4	5
13. Situation entails conflict, disagreement, or argument (example: co-workers disagreeing about the best way to solve a client problem).	1	2	3	4	5
14. Situation includes other persons or groups who count on the actor to do something (example: a team counting on the actor to develop a presentation for a meeting).	1	2	3	4	5
15. Situation involves the actor being responsible for a significant proportion of a task, where the results of the task can be clearly identified by others (example: the actor is writing a client report).	1	2	3	4	5
16. Situation requires a high degree of intellectual or emotional effort.	1	2	3	4	5
17. Situation involves other people who are available to provide reassurance.	1	2	3	4	5
18. Situation might evoke warmth or compassion.	1	2	3	4	5
19. Situation provides opportunities for advice and assistance from others (example: a mentor or superior is available to provide guidance on a task).	1	2	3	4	5
20. Situation requires cooperation among persons or groups in order to reach success.	1	2	3	4	5

Table 7 (cont'd)

21. Situation entails reliance on other people or groups for task completion (example: situation involves team members working on different pieces of a client project).	1	2	3	4	5
22. Situation requires interacting with or assisting persons or groups external to the organization (example: the actor is running a public relations project with community volunteers).	1	2	3	4	5
23. Situation raises moral or ethical issues or affords an opportunity to demonstrate integrity (example: actor learns that a friend is stealing).	1	2	3	4	5
24. Situation permits the opportunity for the actor to deceive someone else (example: actor is attempting to sell a product to a potential client who knows little about the product).	1	2	3	4	5
25. Situation contains another person or group who is being deceptive (example: a team member is attempting to deceive others on the team so that she/he doesn't have to do as much work).	1	2	3	4	5
26. Situation involves interaction between individuals who differ in power or status (example: a student is interacting with a professor).	1	2	3	4	5
27. Situation involves issues of power (example: a supervisor distributing rewards to subordinates).	1	2	3	4	5
28. Situation allows for one's emotions to be freely expressed.	1	2	3	4	5

Table 7 (cont'd)

29. Situation includes behavioral limits (example: rules or social norms that might or might not be challenged).	1	2	3	4	5
30. Situation involves frustration (example: a goal is being blocked) or stress.	1	2	3	4	5
31. Situation includes events that would leave most people feeling negative (examples: angry, anxious, sad).	1	2	3	4	5
32. Situation contains events that would make most people feel positive (examples: happy, accomplished, respected).	1	2	3	4	5
33. Situation involves matters of fairness or justice (example: a person feels as if they've been treated unfairly).	1	2	3	4	5
34. Situation includes people who are diverse (examples: different ethnic, cultural, or religious backgrounds; a variety of perspectives or opinions).	1	2	3	4	5
35. Situation calls for convincing or persuading someone else of something (example: trying to persuade a supervisor for an extension on a deadline).	1	2	3	4	5
36. Situation is one in which the actor controls resources needed by others (example: actor has expertise in an area required to complete a project).	1	2	3	4	5

Table 7 (cont'd)

37. Situation is one in which the actor is the primary focus of attention or is being evaluated by others (example: actor is receiving a performance review).	1	2	3	4	5
38. Situation involves blame or criticism being directed toward the actor (example: actor is being blamed for a project done poorly).	1	2	3	4	5
39. Situation involves other persons or groups that might have conflicting or hidden motives (example: actor is asked to work a weekend when she/he already had plans).	1	2	3	4	5
40. Situation involves task and social demands that are familiar to most people.	1	2	3	4	5
41. Situation is similar to the types of situations that the average person has dealt with.	1	2	3	4	5
42. Situation entails demands that could be managed by the average person.	1	2	3	4	5
43. Situation is one in which the approval of others would likely affect how one would respond or behave.	1	2	3	4	5

Table 8.

Ten-Item Personality Inventory (TIPI) Item Content.

Here are a number of personality traits that may or may not apply to the behaviors presented after each situation. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to each behavior, even if one characteristic applies more strongly than the other.

1. _____ Extraverted, enthusiastic.
2. _____ Critical, quarrelsome.
3. _____ Dependable, self-disciplined.
4. _____ Anxious, easily upset.
5. _____ Open to new experiences, complex.
6. _____ Reserved, quiet.
7. _____ Sympathetic, warm.
8. _____ Disorganized, careless.
9. _____ Calm, emotionally stable.
10. _____ Conventional, uncreative.

Adapted from Gosling, Rentfrow, & Swann (2003)

Table 9.

Descriptive Statistics for ρ Values across Situational Characteristic Ratings.

Rater	*M*	*Median*	*SD*	*Min*	*Max*
1	0.37	0.48	0.30	0.00	0.89
2	0.47	0.65	0.34	0.00	0.94
3	0.53	0.73	0.36	0.00	0.96
4	0.56	0.79	0.37	0.00	0.97
5	0.58	0.82	0.38	0.00	0.97
6	0.60	0.85	0.38	0.00	0.98
7	0.62	0.87	0.38	0.00	0.98
8	0.63	0.88	0.39	0.00	0.98
9	0.64	0.89	0.39	0.00	0.99
10	0.65	0.90	0.39	0.00	0.99
11	0.65	0.91	0.39	0.00	0.99
12	0.66	0.92	0.39	0.00	0.99
13	0.67	0.92	0.39	0.00	0.99
14	0.67	0.93	0.39	0.00	0.99
15	0.68	0.93	0.39	0.00	0.99
16	0.68	0.94	0.39	0.00	0.99
17	0.69	0.94	0.39	0.00	0.99
18	0.69	0.94	0.39	0.00	0.99
19	0.69	0.95	0.39	0.00	0.99
20	0.70	0.95	0.39	0.00	0.99

Table 10.

Recomputed Descriptive Statistics for ρ Values across Situational Characteristic Ratings.

Rater	*M*	*Median*	*SD*	*Min*	*Max*
1	0.40	0.51	0.29	0.00	0.89
2	0.51	0.68	0.33	0.00	0.94
3	0.57	0.76	0.34	0.00	0.96
4	0.60	0.81	0.35	0.00	0.97
5	0.63	0.84	0.35	0.00	0.97
6	0.65	0.86	0.36	0.00	0.98
7	0.66	0.88	0.36	0.00	0.98
8	0.68	0.89	0.36	0.00	0.98
9	0.69	0.91	0.36	0.00	0.99
10	0.70	0.91	0.36	0.00	0.99
11	0.70	0.92	0.36	0.00	0.99
12	0.71	0.93	0.36	0.00	0.99
13	0.72	0.93	0.36	0.00	0.99
14	0.72	0.94	0.36	0.00	0.99
15	0.73	0.94	0.35	0.00	0.99
16	0.73	0.94	0.35	0.00	0.99
17	0.74	0.95	0.35	0.00	0.99
18	0.74	0.95	0.35	0.00	0.99
19	0.75	0.95	0.35	0.00	0.99
20	0.75	0.95	0.35	0.00	0.99

Table 11.

Descriptive Statistics for ρ Values across Behavioral Characteristic Ratings.

Rater	*M*	*Median*	*SD*	*Min*	*Max*
1	0.16	0.15	0.14	0.00	0.48
2	0.42	0.44	0.25	0.00	0.84
3	0.49	0.53	0.26	0.00	0.89
4	0.53	0.59	0.26	0.00	0.91
5	0.56	0.64	0.26	0.00	0.93
6	0.59	0.67	0.27	0.00	0.94
7	0.61	0.70	0.27	0.00	0.95
8	0.63	0.73	0.27	0.00	0.96
9	0.64	0.74	0.27	0.00	0.96
10	0.65	0.75	0.27	0.00	0.96
11	0.66	0.76	0.27	0.00	0.97
12	0.67	0.76	0.27	0.00	0.97
13	0.68	0.77	0.27	0.00	0.97
14	0.69	0.77	0.27	0.00	0.97
15	0.69	0.78	0.27	0.00	0.98
16	0.70	0.78	0.27	0.00	0.98
17	0.70	0.79	0.27	0.00	0.98
18	0.71	0.79	0.27	0.00	0.98
19	0.71	0.79	0.28	0.00	0.98
20	0.72	0.79	0.28	0.00	0.98

Table 12.

Descriptions of and Items Associated with the Situational Characteristic Inventory Scales.

Scale	Items
Task Demands: Situational features associated with the performance of tasks or with outcomes of task performance. These features do not directly involve others (e.g., co-workers, supervisors, subordinates, clients, etc.).	1, 3, 5, 6, 8, 16
Competition/Power: Situational features associated with competition and power (control over status, resources, etc.) within and between groups in accomplishing goals. The emphasis here is on characteristics associated with resource sharing, interdependence, etc. in accomplishing tasks, as opposed to interpersonal treatment or individual behavior.	9, 26, 27, 36, 39
Interpersonal Relations: Situational features associated with positive or negative interactions among people in a situation. The feature might be present in the context of a group task (though it need not be); the emphasis here, however, is on individual behavior and interpersonal treatment.	11, 12, 13, 17, 19, 25
Morality & Fairness: Situational features associated with issues pertaining to morality, ethics, integrity, and fair treatment of individuals.	23, 24, 33
Individual Emotional: Situational features associated with the actor's emotions and management of her/his emotions in a situation.	2, 18, 30, 31, 32
Familiarity/Difficulty: The extent to which features inherent in the situation are (1) likely to be familiar to most people and (2) difficult to manage or navigate.	40, 41, 42
Social Pressure/Social Performance: Situational features associated with being evaluated in a social settings and having to perform in a social manner (e.g., convincing others).	4, 15, 35, 37, 38
Team Task Work: Situational features associated with task performance conducted with the context of a team environment and, hence, includes features relevant to performance in team settings.	10, 14, 20, 21

Note. Content associated with scale items shown in Appendix A.

Table 13.

Situational Characteristic Ratings: Descriptive Statistics and Estimated ρ Values and Variance Components.

						Variance Component Estimate		
	Mean	*SD*	*n*	ICC(2, 1)	ICC(2, *k*)	Stem	Rater	Residual
SC 1	4.04	1.18	934	.181	.693	0.253	0.285	0.859
SC 2	3.96	1.18	928	.200	.717	0.282	0.253	0.875
SC 3	3.59	1.30	932	.249	.772	0.425	0.272	1.009
SC 4	4.14	1.11	933	.243	.766	0.294	0.161	0.756
SC 5	3.52	1.34	930	.210	.730	0.376	0.266	1.148
SC 6	3.33	1.35	929	.177	.685	0.325	0.420	1.096
SC 7	3.49	1.36	926	.199	.715	0.365	0.228	1.244
SC 8	3.01	1.48	928	.304	.816	0.670	0.218	1.313
SC 9	2.75	1.47	933	.193	.710	0.418	0.446	1.301
SC 10	3.16	1.46	933	.327	.832	0.696	0.296	1.136
SC 11	2.50	1.36	930	.294	.809	0.556	0.374	0.959
SC 12	2.91	1.42	928	.313	.822	0.636	0.318	1.080
SC 13	3.35	1.42	926	.311	.820	0.622	0.236	1.142
SC 14	3.32	1.43	932	.276	.796	0.555	0.218	1.237
SC 15	3.41	1.40	929	.214	.735	0.421	0.275	1.271
SC 16	3.52	1.21	932	.107	.551	0.157	0.318	0.990
SC 17	3.19	1.23	924	.062	.401	0.095	0.387	1.042
SC 18	2.42	1.28	928	.095	.517	0.157	0.491	1.000
SC 19	3.53	1.29	930	.129	.601	0.216	0.326	1.133
SC 20	3.76	1.35	934	.282	.800	0.504	0.212	1.073
SC 21	3.25	1.44	927	.298	.811	0.614	0.200	1.247
SC 22	2.84	1.42	925	.116	.571	0.237	0.431	1.367
SC 23	2.88	1.50	933	.266	.787	0.595	0.410	1.231
SC 24	2.37	1.34	931	.140	.624	0.254	0.477	1.079
SC 25	2.38	1.38	930	.125	.593	0.240	0.474	1.203
SC 26	3.23	1.46	932	.224	.747	0.474	0.385	1.252
SC 27	2.89	1.44	929	.173	.680	0.353	0.406	1.277
SC 28	3.15	1.25	932	.102	.537	0.157	0.423	0.953
SC 29	3.45	1.26	932	.117	.575	0.187	0.502	0.908
SC 30	4.22	1.00	932	.195	.712	0.196	0.189	0.617
SC 31	3.87	1.15	927	.188	.700	0.247	0.266	0.805
SC 32	2.45	1.29	929	.159	.657	0.263	0.463	0.932

Table 13 (cont'd)

SC 33	3.20	1.38	929	.274	.793	0.521	0.284	1.098
SC 34	2.70	1.40	928	.218	.739	0.426	0.517	1.010
SC 35	3.30	1.40	929	.193	.709	0.374	0.363	1.199
SC 36	2.87	1.37	927	.107	.547	0.199	0.330	1.334
SC 37	3.05	1.41	927	.107	.549	0.214	0.443	1.338
SC 38	2.83	1.42	932	.125	.594	0.254	0.565	1.205
SC 39	2.72	1.37	932	.140	.624	0.265	0.325	1.299
SC 40	3.62	1.14	929	.043	.315	0.056	0.443	0.794
SC 41	3.70	1.13	926	.060	.392	0.076	0.480	0.709
SC 42	3.73	1.11	927	.056	.375	0.068	0.467	0.685
SC 43	3.64	1.22	929	.087	.491	0.130	0.362	1.004

Table 14.

Situational Characteristic Composite Scores: Descriptive Statistics and Estimated ρ Values and Variance Components.

						Variance Component Estimate		
	Mean	*SD*	*n*	ICC(2, 1)	ICC(2, *k*)	Stem	Rater	Residual
Task Demands	3.50	0.87	934	.276	.796	0.209	0.209	0.337
Competition	2.89	0.96	934	.215	.737	0.196	0.293	0.422
Interpersonal Relations	2.98	0.85	934	.231	.754	0.170	0.238	0.328
Morality & Fairness	2.82	1.07	934	.288	.806	0.332	0.306	0.513
Individual Emotional	3.39	0.62	934	.121	.585	0.046	0.142	0.195
Familiarity & Difficulty	3.68	1.00	934	.060	.394	0.060	0.439	0.501
Social Pressure & Performance	3.35	0.85	934	.205	.725	0.147	0.240	0.332
Team Task Work	3.37	1.13	934	.417	.880	0.522	0.180	0.550

Table 15.

Likelihood Ratio Tests of Between-Stem Variability in Situational Characteristic Ratings.

	No Variance in Rating across Stems			Variance in Rating across SJT Stems			Model Comparison	
	AIC	*BIC*	*-2LL*	*AIC*	*BIC*	*-2LL*	χ^2	*p*
SC 1	2912.07	2926.58	-1453.03	2812.20	2831.56	-1402.10	101.86	.000
SC 2	2910.96	2925.46	-1452.48	2803.43	2822.76	-1397.72	109.53	.000
SC 3	3123.05	3137.56	-1558.52	2958.87	2978.21	-1475.43	166.18	.000
SC 4	2810.51	2825.02	-1402.25	2664.88	2684.24	-1328.44	147.62	.000
SC 5	3156.64	3171.15	-1575.32	3041.10	3060.44	-1516.55	117.54	.000
SC 6	3133.31	3147.81	-1563.65	3041.42	3060.76	-1516.71	93.88	.000
SC 7	3178.44	3192.93	-1586.22	3074.28	3093.60	-1533.14	106.16	.000
SC 8	3353.85	3368.35	-1673.93	3161.38	3180.71	-1576.69	194.48	.000
SC 9	3316.96	3331.48	-1655.48	3206.06	3225.41	-1599.03	112.90	.000
SC 10	3326.70	3341.21	-1660.35	3096.31	3115.66	-1544.15	232.39	.000
SC 11	3179.59	3194.09	-1586.79	2968.37	2987.71	-1480.19	213.21	.000
SC 12	3260.02	3274.52	-1627.01	3042.84	3062.17	-1517.42	219.18	.000
SC 13	3260.98	3275.47	-1627.49	3048.56	3067.88	-1520.28	214.42	.000
SC 14	3269.23	3283.74	-1631.61	3114.40	3133.75	-1553.20	156.83	.000
SC 15	3241.74	3256.25	-1617.87	3126.25	3145.58	-1559.12	117.50	.000
SC 16	2944.58	2959.09	-1469.29	2900.09	2919.44	-1446.04	46.49	.000
SC 17	2931.21	2945.69	-1462.60	2914.36	2933.68	-1453.18	18.84	.000
SC 18	2985.40	2999.90	-1489.70	2946.86	2966.19	-1469.43	40.54	.000
SC 19	3074.31	3088.82	-1534.16	3017.04	3036.38	-1504.52	59.27	.000
SC 20	3171.87	3186.39	-1582.93	3000.95	3020.31	-1496.47	172.92	.000
SC 21	3295.37	3309.86	-1644.68	3104.99	3124.32	-1548.50	192.37	.000
SC 22	3225.54	3240.03	-1609.77	3179.54	3198.86	-1585.77	48.00	.000

Table 15 (cont'd)

SC 23	3361.33	3375.85	-1677.67	3179.46	3198.82	-1585.73	183.87	.000
SC 24	3112.00	3126.50	-1553.00	3035.44	3054.79	-1513.72	78.55	.000
SC 25	3171.58	3186.08	-1582.79	3110.59	3129.93	-1551.29	62.99	.000
SC 26	3284.13	3298.64	-1639.06	3166.26	3185.61	-1579.13	119.87	.000
SC 27	3236.44	3250.94	-1615.22	3157.78	3177.12	-1574.89	80.65	.000
SC 28	2938.76	2953.27	-1466.38	2904.37	2923.72	-1448.19	36.38	.000
SC 29	2959.58	2974.09	-1476.79	2899.41	2918.76	-1445.71	62.17	.000
SC 30	2608.48	2622.99	-1301.24	2494.45	2513.80	-1243.22	116.03	.000
SC 31	2828.78	2843.28	-1411.39	2733.7	2753.02	-1362.85	97.09	.000
SC 32	3009.24	3023.74	-1501.62	2918.94	2938.28	-1455.47	92.30	.000
SC 33	3215.13	3229.63	-1604.56	3032.66	3052.00	-1512.33	184.47	.000
SC 34	3164.82	3179.32	-1579.41	3020.6	3039.93	-1506.30	146.22	.000
SC 35	3204.07	3218.58	-1599.04	3101.76	3121.10	-1546.88	104.31	.000
SC 36	3166.13	3180.62	-1580.06	3130.2	3149.52	-1561.10	37.93	.000
SC 37	3209.56	3224.06	-1601.78	3167.35	3186.68	-1579.68	44.21	.000
SC 38	3201.07	3215.58	-1597.54	3142.84	3162.19	-1567.42	60.23	.000
SC 39	3202.55	3217.07	-1598.28	3140.24	3159.59	-1566.12	64.31	.000
SC 40	2727.80	2742.31	-1360.90	2719.22	2738.56	-1355.61	10.58	.001
SC 41	2664.32	2678.82	-1329.16	2647.67	2667.00	-1319.84	18.65	.000
SC 42	2633.67	2648.16	-1313.83	2617.75	2637.08	-1304.88	17.91	.000
SC 43	2937.32	2951.82	-1465.66	2906.84	2926.18	-1449.42	32.48	.000

Table 16.

Likelihood Ratio Tests of Between-Stem Variability in Situational Characteristic Composites.

	No Variance in Rating across Stems			Variance in Rating across SJT Stems			Model Comparison	
	AIC	*BIC*	*-2LL*	*AIC*	*BIC*	*-2LL*	χ^2	*p*
Task Demands	2306.91	2321.43	-1150.45	2067.13	2086.49	-1029.57	241.77	.000
Competition	2433.90	2448.42	-1213.95	2271.62	2290.98	-1131.81	164.28	.000
Interpersonal Relations	2238.87	2253.39	-1116.43	2048.51	2067.87	-1020.25	192.36	.000
Morality & Fairness	2705.86	2720.38	-1349.93	2456.66	2476.02	-1224.33	251.20	.000
Individual Emotional	1589.27	1603.79	-791.64	1515.47	1534.83	-753.74	75.80	.000
Familiarity & Difficulty	2409.55	2424.07	-1201.77	2388.06	2407.42	-1190.03	23.49	.000
Social Pressure & Performance	2205.37	2219.89	-1099.68	2049.90	2069.25	-1020.95	157.47	.000
Team Task Work	2835.53	2850.05	-1414.77	2479.68	2499.04	-1235.84	357.85	.000

Table 17.

Behavioral Characteristic Ratings: Descriptive Statistics and Estimated ρ Values and Variance Components.

						Variance Component Estimate				
	Mean	*SD*	*n*	ICC(3, 1)	ICC(3, *k*)	Response Option	Stem	Rater	Stem-Rater Interaction	Residual
Agreeableness Neg.	3.86	1.84	5,856	.150	.878	0.511	0.164	0.483	0.249	2.000
Agreeableness Pos.	4.22	1.55	5,853	.227	.931	0.545	0.070	0.222	0.194	1.371
Conscientiousness Neg.	3.05	1.63	5,851	.182	.902	0.482	0.016	0.408	0.144	1.591
Conscientiousness Pos.	4.85	1.63	5,836	.249	.958	0.653	0.000	0.225	0.065	1.680
Emotional Stability Neg.	3.61	1.63	5,830	.144	.865	0.384	0.067	0.371	0.251	1.598
Emotional Stability Pos.	4.72	1.48	5,853	.162	.895	0.358	0.034	0.307	0.131	1.380
Extraversion Neg.	3.41	1.65	5,849	.229	.933	0.622	0.000	0.335	0.130	1.634
Extraversion Pos.	4.70	1.79	5,867	.263	.949	0.841	0.090	0.321	0.141	1.804
Openness Neg.	3.67	1.54	5,861	.080	.783	0.192	0.000	0.437	0.127	1.632
Openness Pos.	4.44	1.63	5,847	.202	.949	0.538	0.163	0.135	0.140	1.683

Table 18.

Behavioral Characteristic Composite Scores: Descriptive Statistics and Estimated ρ Values and Variance Components.

						Variance Component Estimate				
	Mean	*SD*	*n*	ICC(3, 1)	ICC(3, *k*)	Response Option	Stem	Rater	Stem-Rater Interaction	Residual
Agreeableness	4.18	1.40	5,877	.240	.938	0.471	0.083	0.184	0.130	1.093
Conscientiousness	4.65	1.40	5,877	.315	.963	0.613	0.000	0.164	0.078	1.093
Emotional Stability	4.90	1.42	5,877	.273	.951	0.547	0.000	0.233	0.064	1.159
Extraversion	4.55	1.32	5,875	.196	.911	0.346	0.041	0.226	0.127	1.022
Openness	4.38	1.31	5,877	.193	.931	0.327	0.030	0.175	0.075	1.090

Table 19.

Likelihood Ratio Tests of Between-Option Variability in Behavioral Characteristic Ratings.

	No Variance in Rating across Response Options			Variance in Rating across SJT Response Options			Model Comparison	
	AIC	*BIC*	*-2LL*	*AIC*	*BIC*	*-2LL*	χ^2	*p*
Agreeableness Neg.	22420.24	22446.94	-11206.12	22287.34	22327.40	-11137.67	136.90	.000
Agreeableness Pos.	20329.03	20355.73	-10160.51	20195.47	20235.52	-10091.74	137.56	.000
Conscientiousness Neg.	20921.75	20948.44	-10456.87	20859.55	20899.59	-10423.77	66.20	.000
Conscientiousness Pos.	20954.78	20981.47	-10473.39	20944.70	20984.74	-10466.35	14.07	.001
Emotional Stability Neg.	21080.51	21107.19	-10536.25	20907.61	20947.64	-10447.81	176.89	.000
Emotional Stability Pos.	20046.34	20073.04	-10019.17	19975.54	20015.58	-9981.77	74.81	.000
Extraversion Neg.	21066.81	21093.50	-10529.40	21023.28	21063.32	-10505.64	47.53	.000
Extraversion Pos.	21809.75	21836.46	-10900.88	21756.67	21796.74	-10872.34	57.08	.000
Openness Neg.	20726.94	20753.64	-10359.47	20681.34	20721.40	-10334.67	49.60	.000
Openness Pos.	21147.77	21174.46	-10569.88	21063.76	21103.81	-10525.88	88.00	.000

Note. “Neg” denotes a negative-keyed item for each dimension. “Pos” denotes a positively-keyed item for each dimension.

Table 20.

Likelihood Ratio Tests of Between-Stem Variability in Behavioral Characteristic Composites.

	No Variance in Rating across Response Options			Variance in Rating across SJT Response Options			Model Comparison	
	AIC	*BIC*	*-2LL*	*AIC*	*BIC*	*-2LL*	χ^2	*p*
Agreeableness	19048.59	19075.31	-9520.30	18938.80	18978.87	-9463.40	113.79	.000
Conscientiousness	19134.63	19161.34	-9563.31	19110.75	19150.82	-9549.37	27.88	.000
Emotional Stability	18595.71	18622.42	-9293.85	18475.30	18515.37	-9231.65	124.40	.000
Extraversion	18890.88	18917.59	-9441.44	18852.50	18892.58	-9420.25	42.37	.000
Openness	18638.58	18665.29	-9315.29	18597.13	18637.20	-9292.56	45.45	.000

Table 21.

Descriptive Statistics: Situational Characteristic Composites.

	Mean	*SD*	1	2	3	4	5	6	7	8
1. Task Demands	3.51	0.50	.811							
2. Competition and Power	2.91	0.52	.386	.768						
3. Interpersonal Relationships	2.99	0.46	.031	.459	.797					
4. Moral Issues and Fairness	2.83	0.63	-.125	.485	.686	.761				
5. Individual-Emotional	3.38	0.26	.129	.113	.396	.273	.838			
6. Familiarity and Difficulty	3.67	0.44	-.282	-.404	-.221	-.225	.067	.918		
7. Social Pressure & Performance	3.36	0.45	.536	.498	.421	.178	.245	-.197	.683	
8. Team Task Work	3.39	0.77	.493	.511	.316	.092	.012	-.295	.595	.922

Note. Sample size for all estimates is $n = 90$. Reliability estimates (Cronbach's α) shown along the matrix diagonal. Correlations equal to or greater than 0.207 in absolute magnitude significant at the $p < 0.05$ level.

Table 22.

Descriptive Statistics: FFM Trait Expression (Behavioral Characteristic) Composites.

	Mean	*SD*	1	2	3	4	5
1. Agreeableness	4.18	0.82	.735				
2. Conscientiousness	4.91	0.83	.446	.883			
3. Emotional Stability	4.56	0.71	.765	.588	.808		
4. Extraversion	4.66	0.85	.196	.518	.350	.657	
5. Openness	4.39	0.70	.471	.576	.515	.677	.684

Note. Sample size for all estimates is n = 534. Reliability estimates (Cronbach's α) shown along the matrix diagonal. Correlations equal to or greater than 0.207 significant at the $p < 0.05$ level.

Table 23.

Descriptive Statistics: Stem-Level Individual Difference Correlations (r-to-)z Values.

	Mean	*SD*	*n*	1	2	3	4	5	6	7	8
1. Ability	.063	.088	89	–							
2. Agreeableness	.127	.075	89	-.008	–						
3. Conscientiousness	.100	.074	89	-.135	.476	–					
4. Emotional Stability	.032	.059	89	.205	.295	.244	–				
5. Extraversion	.031	.082	89	-.051	.358	.164	.322	–			
6. Openness to Experience	.080	.067	89	.135	.488	.247	.214	.320	–		
7. Experience (Business Courses)	.027	.059	43	.072	-.057	-.053	-.131	-.221	-.036	–	
8. Experience (Job Tenure)	.006	.073	43	-.007	.052	.268	-.033	.027	.061	.390	–

Note. For ability and FFM characteristics, $|r| \geq .208$ significant at $p < .05$. For experience, $|r| \geq .301$ significant at $p < .05$. The sample size is lower for the experience variables as data for these variables were available for only a subset of the total sample of stems (the M-SJI).

Table 24.

Descriptive Statistics: Stem-Level Criterion Correlations.

	Mean	*SD*	*n*	1	2	3	4	5	6	7
1. Deviance: Time 1	-.062	.053	36	–						
2. Deviance: Time 2	-.071	.062	36	.683	–					
3. GPA	.046	.061	79	-.323	-.337	–				
4. OCB: Time 1	.045	.061	36	-.319	-.063	.163	–			
5. OCB: Time 2	.032	.077	36	.000	.180	-.023	.776	–		
6. BARS: Time 1	.118	.065	36	-.509	-.392	.337	.683	.573	–	
7. BARS: Time 2	.090	.060	36	-.294	-.248	.096	.615	.710	.736	–

Note. For GPA, $|r| \geq .221$ significant at $p < .05$. For all other correlations, $|r| \geq .329$ significant at $p < .05$. The sample size is larger for GPA as data for this variable were available for both the CB-SJI and M-SJI; otherwise, all outcome data is for the CB-SJI.

Table 25.

Descriptive Statistics: Response Option-Level Individual Difference Correlations.

	Mean	*SD*	*n*	1	2	3	4	5	6	7	8
1. Ability	-.014	.075	534	–							
2. Agreeableness	-.009	.090	534	.155	–						
3. Conscientiousness	-.005	.079	534	.016	.652	–					
4. Emotional Stability	-.007	.063	534	.157	.349	.382	–				
5. Extraversion	-.002	.072	534	-.015	.397	.297	.376	–			
6. Openness	-.008	.077	534	.221	.636	.470	.368	.378	–		
7. Experience (Business Courses)	-.003	.072	246	.164	.098	.014	.015	.103	.107	–	
8. Experience (Job Tenure)	-.004	.073	246	.128	.071	.135	.058	.137	.145	.399	–

Note. For ability and FFM characteristics, $|r| \geq .085$ significant at $p < .05$. For Experience (Business Courses) and Experience (Job Tenure), $|r| \geq .125$ significant at $p < .05$.

Table 26.

Descriptive Statistics: Response Option-Level Criterion Correlations.

	Mean	*SD*	*n*	1	2	3	4	5	6	7
1. Deviance: Time 1	.008	.060	188	–						
2. Deviance: Time 2	.008	.066	188	.633	–					
3. GPA	-.005	.064	434	-.268	-.287	–				
4. OCB: Time 1	-.005	.065	188	-.217	-.232	.039	–			
5. OCB: Time 2	-.003	.069	188	-.136	-.075	.042	.781	–		
6. BARS: Time 1	-.011	.079	188	-.560	-.485	.323	.613	.515	–	
7. BARS: Time 2	-.011	.075	188	-.473	-.491	.211	.677	.658	.793	–

Table 27.

Correlations between Situational Characteristic Composite Scores and Individual Difference Correlations.

	Task Demands	Competition	Interpersonal Relations	Morality & Fairness	Individual Emotional
Ability	-.015	**.301**	**.241**	.160	-.012
Agreeableness	-.098	-.082	.091	.067	.050
Conscientiousness	.124	.090	.028	-.045	-.048
Emotional Stability	.044	.011	.028	-.081	-.024
Extraversion	.164	.109	-.175	-.**242**	-.139
Openness to Experience	-.025	.059	.024	-.136	-.193
Experience (Business Courses)	.222	.013	.073	-.197	.094
Experience (Job Tenure)	.031	-.135	.171	-.072	.080

Note. Estimates associated with Ability and FFM correlations based on a sample size of $n = 89$ stems; for these estimates, $|r| \geq .208$ significant at $p < .05$. Estimates associated with Experience (Business Courses) and Experience (Job Tenure) correlations based on a sample size of $n = 43$ stems; for these estimates, $|r| \geq .301$ significant at $p < .05$ (highlighted in bold typeface).

Table 27 (cont'd)

Familiarity & Difficulty	Social Pressure & Performance	Team Task Work
-.190	.179	.184
.032	-.153	-.078
-.106	.062	.081
-.177	.072	.010
-.185	-.070	.127
-.128	-.087	.117
.280	.222	.083
.131	.136	-.061

Table 28.

Correlations between Situational Characteristic Composite Scores and Criterion Correlations.

	Deviance: Time 1	Deviance: Time 2	GPA	OCB: Time 1	OCB: Time 2	BARS Time 1	BARS Time 2
Task Demands	.030	-.071	.095	.092	.083	.099	.103
Competition	-.188	-.291	.182	-.057	-.293	.070	-.137
Interpersonal Relations	**-.334**	**-.509**	.142	-.018	-.296	.047	-.115
Morality & Fairness	**-.484**	**-.467**	**.230**	.070	-.246	.129	-.033
Individual Emotional	-.173	-.224	-.009	.134	.006	.107	.152
Familiarity & Difficulty	.134	.278	-.041	**.384**	**.351**	.192	.188
Social Pressure & Performance	-.115	-.321	.092	-.013	-.218	-.013	-.040
Team Task Work	-.033	-.261	-.015	-.218	-.278	-.052	-.145

Note. Estimates associated with GPA correlations based on a sample size of $n = 79$ stems; for these estimates, $|r| \geq .221$ significant at $p < .05$. For all other criterion correlations, estimates based on a sample size of $n = 36$ stems; for these estimates, $|r| \geq .329$ significant at $p < .05$ (highlighted in bold typeface).

Table 29.

Correlations between Behavioral Characteristic Composite Scores and Individual Difference Correlations.

	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness
Ability	**.150**	**.122**	**.158**	**.124**	**.144**
Agreeableness	**.198**	**.444**	**.298**	**.313**	**.288**
Conscientiousness	**.148**	**.453**	**.218**	**.302**	**.255**
Emotional Stability	.075	**.239**	**.149**	**.248**	**.198**
Extraversion	-.049	**.103**	.037	**.236**	**.150**
Openness	.087	**.291**	**.166**	**.297**	**.284**
Experience (Business Courses)	.022	.112	.061	**.149**	.115
Experience (Job Tenure)	**.170**	**.149**	**.133**	**.217**	**.249**

Note. Response option correlations denote the variables comprising the rows of the matrix; behavioral characteristic ratings denote the variables comprising the columns of the matrix. Estimates associated with ability and FFM response option correlations based on a sample size of $n = 534$ response options; for these estimates, $|r| \geq .085$ significant at $p < .05$. Estimates associated with Experience (Business Courses) and Experience (Job Tenure) response option correlations based on a sample size of $n = 246$ response options; for these estimates, $|r| \geq .125$ significant at $p < .05$ (highlighted in bold typeface).

Table 30.

Correlations between Behavioral Characteristic Composite Scores and Criterion Correlations.

	Deviance: Time 1	Deviance: Time 2	GPA	OCB: Time 1	OCB: Time 2	BARS: Time 1	BARS: Time 2
Agreeableness	**-.178**	**-.210**	**.208**	**.153**	.083	**.266**	**.190**
Conscientiousness	**-.473**	**-.470**	**.210**	**.283**	**.221**	**.585**	**.488**
Emotional Stability	**-.198**	**-.202**	**.197**	.139	.046	**.287**	**.197**
Extraversion	-.119	-.020	**.117**	**.355**	**.349**	**.428**	**.363**
Openness	**-.180**	-.142	**.108**	**.351**	**.314**	**.458**	**.386**

Note. Response option correlations denote the variables comprising the columns of the matrix; behavioral characteristic ratings denote the variables comprising the rows of the matrix. Estimates associated with GPA response option correlations based on a sample size of $n = 434$ response options; for these estimates, $|r| \geq .094$ significant at $p < .05$. All other estimates based on a sample size of $n = 188$ response options; for these estimates, $|r| \geq .143$ significant at $p < .05$ (highlighted in bold typeface).

Table 31.

Model Estimates: Individual Difference Correlations Regressed on Situational Characteristic Composite Scores.

	Ability	Agreeableness	Conscientiousness	Emotional Stability
Intercept	.063 (.009)*	.127 (.008)*	.100 (.008)*	.032 (.006)*
Task Demands	-.042 (.025)	.008 (.022)	.014 (.022)	-.006 (.018)
Competition	.051 (.026)*	-.014 (.023)	.012 (.023)	-.001 (.018)
Interpersonal Relations	.033 (.032)	.035 (.028)	.019 (.028)	.018 (.022)
Morality & Fairness	-.022 (.023)	.002 (.020)	-.016 (.020)	-.021 (.016)
Individual Emotional	-.018 (.040)	.008 (.035)	-.018 (.036)	-.005 (.028)
Familiarity & Difficulty	-.021 (.025)	.001 (.022)	-.010 (.022)	-.030 (.017)*
Social Pressure & Performance	.017 (.030)	-.041 (.026)	-.006 (.027)	.015 (.021)
Team Task Work	.003 (.017)	.003 (.015)	-.003 (.015)	-.009 (.012)
Model R	.387	.247	.184	.257
Model R^2	.150	.061	.034	.066
Model F	$F(8, 80) = 1.767$	$F(8, 80) = 0.648$	$F(8, 80) = 0.350$	$F(8, 80) = 0.701$

Note. Parameter estimate (standard error). Estimates significant at the $p < .05$ level denoted by an asterisk.

Table 31 (cont'd)

Extraversion	Openness to Experience	Experience (Business Courses)	Experience (Work Tenure)
.031 (.008)*	.080 (.007)*	.029 (.010)*	.009 (.013)
.010 (.023)	-.012 (.019)	.051 (.030)	.026 (.040)
.044 (.024)*	.022 (.020)	.014 (.026)	-.020 (.034)
-.003 (.029)	.047 (.024)*	.034 (.033)	.046 (.044)
-.045 (.021)*	-.042 (.017)*	-.017 (.029)	-.007 (.038)
.000 (.036)	-.039 (.030)	-.003 (.047)	-.005 (.062)
-.029 (.022)	-.015 (.018)	.057 (.029)*	.027 (.039)
-.048 (.027)*	-.037 (.023)	.005 (.030)	.014 (.039)
.011 (.015)	.011 (.012)	-.017 (.021)	-.014 (.027)
.431	.390	.476	.324
.186	.152	.227	.105
$F(8, 80) = 2.291$	$F(8, 80) = 1.790$	$F(8, 34) = 1.249$	$F(8, 34) = 0.498$

Table 32.

Model Estimates: Criterion Correlations Regressed on Situational Characteristic Composite Scores.

	Deviance: Time 1	Deviance: Time 2	GPA	OCB: Time 1	OCB: Time 2
Intercept	-.059 (.013)*	-.080 (.015)*	.045 (.007)*	.017 (.015)	.000 (.019)
Task Demands	-.008 (.022)	.000 (.026)	.022 (.018)	.028 (.026)	.056 (.034)
Competition	.012 (.035)	.024 (.040)	.011 (.019)	-.011 (.040)	-.050 (.052)
Interpersonal Relations	.023 (.036)	-.035 (.042)	.002 (.024)	.008 (.042)	-.010 (.054)
Morality & Fairness	-.058 (.024)*	-.019 (.029)	.024 (.018)	.025 (.028)	.022 (.037)
Individual Emotional	-.040 (.042)	-.012 (.049)	-.028 (.032)	.019 (.049)	-.009 (.064)
Familiarity & Difficulty	.015 (.029)	.040 (.035)	.013 (.019)	.073 (.034)*	.066 (.045)
Social Pressure & Performance	.011 (.039)	-.033 (.046)	.009 (.023)	-.003 (.045)	-.033 (.059)
Team Task Work	.008 (.016)	.009 (.019)	-.015 (.012)	-.026 (.018)	-.010 (.024)
Model *R*	.563	.561	.317	.546	.517
Model R^2	.317	.314	.101	.299	.267
Model *F*	$F(8, 27) = 1.56$	$F(8, 27) = 1.55$	$F(8, 70) = 0.98$	$F(8, 27) = 1.44$	$F(8, 27) = 1.23$

Table 32 (cont'd)

BARS: Time 1	BARS: Time 2
.102 (.017)*	.075 (.016)*
.022 (.031)	.032 (.028)
.019 (.048)	-.031 (.044)
-.020 (.049)	-.033 (.045)
.034 (.034)	.036 (.031)
.042 (.058)	.033 (.053)
.053 (.041)	.018 (.037)
-.055 (.054)	-.004 (.049)
.001 (.022)	-.009 (.020)
.380	.383
.145	.146
$F(8, 27) = 0.57$	$F(8, 27) = 0.58$

Table 33.

OLS Model Estimates: FFM Personality Correlations Regressed on Behavioral Characteristic Composite Scores.

	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness
Intercept	-.009 (.004)*	-.005 (.003)	-.007 (.003)*	-.002 (.003)	-.008 (.003)*
FFM Predictor	.022 (.005)*	.043 (.004)*	.013 (.004)*	.020 (.004)*	.031 (.005)*
SD: Residual	.088	.070	.062	.070	.074
Model R^2	.039	.205	.022	.056	.081
Model *F*	$F(1, 532) = 21.80$	$F(1, 532) = 137.37$	$F(1, 532) = 12.15$	$F(1, 532) = 31.43$	$F(1, 532) = 46.82$

Note. Standard error provided in parentheses. Estimates significant at $p < .05$ denoted by asterisk.

Table 34.

OLS Slope Estimates for Stems with the Five Most Negative Slopes and Associated Within-Stem, Between-Option Standard Deviations for FFM Trait Expression Ratings.

Agreeableness		Conscientiousness		Emotional Stability		Extraversion		Openness	
b	*SD*	*b*	*SD*	*b*	*SD*	*b*	*SD*	*b*	*SD*
-.247	.437 (-0.623)	-.241	.146 (-1.606)	-.386	.197 (-1.290)	-.288	.191 (-1.675)	-.257	.426 (-0.649)
-.233	.441 (-0.612)	-.215	.252 (-1.310)	-.139	.342 (-0.812)	-.284	.213 (-1.609)	-.201	.300 (-1.116)
-.217	.113 (-1.554)	-.142	.481 (-0.678)	-.129	.396 (-0.632)	-.229	.229 (-1.526)	-.169	.204 (-1.475)
-.163	.278 (-1.079)	-.126	.165 (-1.553)	-.125	.341 (-0.814)	-.213	.714 (-0.116)	-.154	.432 (-0.629)
-.134	.297 (-1.025)	-.091	.222 (-1.394)	-.110	.403 (-0.610)	-.097	.542 (-0.630)	-.139	.367 (-0.870)

Note. Slope estimates correspond to the regression of FFM trait saturation correlations on FFM trait expression ratings within each item stem. *SD* values pertain to the standard deviation of the FFM trait expression ratings within each stem. The value in parentheses corresponds to the *z*-score associated with the *SD* estimate.

Table 35.

Model Comparison: Random-Intercept, Random-Slope Models versus Fixed-Intercept, Random-Slope Models.

	Random-Intercept, Random-Slope (Model 1)			Random-Intercept, Fixed-Slope (Model 2)			Fixed-Intercept, Random-Slope (Model 3)		
	AIC	BIC	-2LL	AIC	BIC	-2LL	AIC	BIC	-2LL
Agreeableness	-1088.29	-1062.61	-1100.29	-1080.44	-1063.32	-1088.44	-1092.29	-1075.17	-1100.29
Conscientiousness	-1370.53	-1344.85	-1382.53	-1344.49	-1327.36	-1352.49	-1374.53	-1357.41	-1382.53
Emotional Stability	-1458.62	-1432.94	-1470.62	-1447.23	-1430.11	-1455.23	-1462.62	-1445.50	-1470.62
Extraversion	-1349.59	-1323.91	-1361.59	-1331.13	-1314.01	-1339.13	-1353.59	-1336.47	-1361.59
Openness	-1278.96	-1253.28	-1290.96	-1277.92	-1260.80	-1285.92	-1282.96	-1265.84	-1290.96

Note. Information criteria and deviance computed for model comparison purposes were estimated using maximum likelihood, as the maximum likelihood deviance (-2LL) is distributed appropriately as a chi-square variate with degrees of freedom equal to the difference in the number of parameters in the two models being compared. Information criteria and deviance show in other tables were generated using restricted maximum likelihood (REML) procedures and, as such, might different slightly from those above.

Table 35 (cont'd)

Model Comparison: 1 versus 2		Model Comparison: 1 versus 3	
$\chi 2(2)$	p	$\chi 2(2)$	p
11.85	.00	.00	1.00
30.05	.00	.00	1.00
15.39	.00	.00	1.00
22.46	.00	.00	1.00
5.04	.08	.00	1.00

Table 36.

Mixed Model Estimates: FFM Personality Correlations Regressed on Behavioral Characteristic Composite Scores.

	Agreeableness	Conscientiousness	Emotional Stability	Extraversion	Openness
Intercept	-.009 (.004)*	-.005 (.003)	-.007 (.003)*	-.002 (.003)	-.008 (.003)*
FFM Predictor	.030 (.007)*	.048 (.006)*	.019 (.006)*	.019 (.006)*	.039 (.006)*
SD: Slopes	.037	.036	.037	.036	.028
SD: Residual	.084	.063	.058	.064	.071
AIC	-1074.90	-1356.13	-1444.24	-1335.20	-1264.95
BIC	-1057.77	-1339.01	-1427.11	-1318.08	-1247.82
-2LL	-1100.30	-1382.53	-1470.62	-1361.59	-1290.96
Model R^2	.168	.404	.195	.255	.191
Number of Response Options	534	534	534	534	534
Number of Stems	90	90	90	90	90

Note. Standard error provided in parentheses. Estimates significant at $p < .05$ denoted by asterisk.

Table 37.

Model Estimates: FFM Personality Correlations Regressed on Behavioral Characteristic and Situational Characteristic Composite Scores.

	Agreeableness	Conscientiousness
Intercept	-.009 (.004)*	-.005 (.003)
FFM Predictor	.035 (.008)*	.048 (.006)*
Task Demands	-.005 (.011)	-.006 (.008)
Competition	-.001 (.011)	-.002 (.008)
Interpersonal Relations	-.003 (.013)	-.003 (.010)
Morality, Integrity, Fairness	.000 (.009)	.000 (.007)
Individual-Emotional	-.004 (.016)	-.005 (.012)
Familiarity, Difficulty	-.005 (.009)	-.007 (.007)
Social Pressure/Performance	.005 (.012)	.005 (.009)
Team Task Work	.002 (.007)	.002 (.005)
FFM Predictor x Task Demands	.004 (.023)	.030 (.016)
FFM Predictor x Competition	-.016 (.020)	.004 (.017)
FFM Predictor x Interpersonal Relations	-.024 (.028)	.055 (.021)*
FFM Predictor x Morality, Integrity, Fairness	-.005 (.020)	-.027 (.014)
FFM Predictor x Individual-Emotional	.013 (.035)	-.054 (.025)*
FFM Predictor x Familiarity, Difficulty	.011 (.020)	-.001 (.015)
FFM Predictor x Social Pressure/Performance	-.041 (.026)	-.041 (.021)*
FFM Predictor x Team Task Work	.019 (.015)	-.005 (.011)
SD: Slopes	.038	.033
SD: Residual	.084	.063
AIC	-945.64	-1224.20
BIC	-860.04	-1138.59
-2LL	-1111.37	-1400.82
Model R^2	.185	.407
Number of Response Options	534	534
Number of Stems	90	90

Note. Standard error provided in parentheses. Estimates significant at $p < .05$ denoted by asterisk.

Table 37 (cont'd)

Emotional Stability	Extraversion	Openness
-.006 (.003)*	-.002 (.003)	-.008 (.003)*
.021 (.007)*	.019 (.006)*	.035 (.006)*
-.005 (.007)	-.001 (.008)	-.003 (.009)
-.001 (.007)	-.004 (.008)	-.005 (.009)
-.003 (.009)	.004 (.010)	-.011 (.011)
.005 (.006)	.003 (.007)	.012 (.008)
-.003 (.011)	-.005 (.012)	-.003 (.013)
.011 (.007)	.007 (.007)	.010 (.008)
.006 (.008)	.000 (.009)	.009 (.010)
.001 (.005)	.000 (.005)	.001 (.006)
-.018 (.019)	.005 (.017)	-.018 (.016)
.009 (.017)	-.009 (.019)	-.014 (.019)
-.004 (.024)	-.002 (.023)	.027 (.021)
-.029 (.016)	.000 (.016)	-.031 (.016)
.006 (.030)	-.019 (.028)	-.024 (.026)
-.032 (.017)	-.021 (.017)	-.033 (.017)*
-.020 (.022)	.001 (.020)	-.002 (.019)
.001 (.012)	-.006 (.011)	.002 (.012)
.036	.038	.018
.059	.065	.071
-1311.55	-1192.91	-1136.91
-1225.94	-1107.30	-1051.30
-1487.18	-1367.85	-1310.54
.211	.266	.183
534	534	534
90	90	90

Table 38.

Mixed Model Estimates: Ability and Experience Correlations Regressed on Behavioral Characteristic Composite Scores.

	SAT/ACT	Experience: Business Courses	Experience: Job Tenure
Agreeableness			
Intercept	-.014 (.003)*	-.003 (.005)	-.004 (.005)
Slope	.019 (.005)*	.003 (.006)	.019 (.006)*
SD: Slopes	.000	.000	.000
Model R^2	.031	.001	.039
Conscientiousness			
Intercept	-.014 (.003)*	-.003 (.005)	-.004 (.005)
Slope	.012 (.004)*	.014 (.007)	.019 (.007)*
SD: Slopes	.000	.000	.000
Model R^2	.015	.015	.039
Emotional Stability			
Intercept	-.014 (.003)*	-.003 (.005)	-.004 (.005)
Slope	.026 (.006)*	.008 (.007)	.016 (.007)*
SD: Slopes	.000	.000	.000
Model R^2	.066	.006	.024
Extraversion			
Intercept	-.014 (.003)*	-.003 (.005)	-.004 (.005)
Slope	.013 (.004)*	.016 (.006)*	.024 (.006)*
SD: Slopes	.000	.000	.000
Model R^2	.017	.026	.056
Openness			
Intercept	-.014 (.003)*	-.003 (.004)	-.004 (.004)
Slope	.021 (.006)*	.016 (.009)	.037 (.008)*
SD: Slopes	.000	.000	.000
Model R^2	.052	.056	.084

Note. Standard error provided in parentheses. Estimates significant at $p < .05$ denoted by asterisk.

Table 39.

OLS Model Estimates: Correlations with Outcome Variables Regressed on Behavioral Characteristic Composite Scores.

	Deviance: Time 1	Deviance: Time 2	GPA
Agreeableness			
Intercept	.009 (.004)*	.011 (.005)*	-.006 (.003)
Slope	-.013 (.005)*	-.017 (.006)*	.016 (.004)*
SD: Residual	.059	.065	.062
Model R^2	.032	.044	.043
Model *F*	6.07	8.61	19.54
Conscientiousness			
Intercept	.008 (.004)*	.009 (.004)*	-.006 (.003)*
Slope	-.030 (.004)*	-.033 (.004)*	.016 (.004)*
SD: Residual	.053	.059	.062
Model R^2	.224	.221	.044
Model *F*	53.65	52.87	19.95
Emotional Stability			
Intercept	.008 (.004)	.008 (.005)	-.005 (.003)
Slope	-.017 (.006)*	-.019 (.007)*	.017 (.004)*
SD: Residual	.059	.065	.063
Model R^2	.039	.041	.039
Model *F*	7.58	7.95	17.48
Extraversion			
Intercept	.007 (.004)	.008 (.005)	-.006 (.003)
Slope	-.008 (.005)	-.002 (.006)	.009 (.004)*
SD: Residual	.060	.067	.063
Model R^2	.014	.000	.014
Model *F*	2.67	0.08	6.01
Openness			
Intercept	.007 (.004)	.008 (.005)	-.006 (.003)
Slope	-.016 (.006)*	-.013 (.007)*	.010 (.004)*
SD: Residual	.059	.066	.063
Model R^2	.033	.020	.012
Model *F*	6.26	3.84	5.05

Note. Standard error provided in parentheses. Estimates significant at $p < .05$ denoted by asterisk. For all outcomes except for GPA, numerator and denominator *df* associated with the *F* statistic are 1 and 186; for GPA, these are 1 and 432.

Table 39 (cont'd)

OCB: Time 1	OCB: Time 2	BARS Total: Time 1	BARS Total: Time 2
-.007 (.005)	-.004 (.005)	-.015 (.006)*	-.013 (.005)*
.012 (.006)*	.007 (.006)	.026 (.007)*	.018 (.007)*
.064	.069	.077	.074
.023	.007	.071	.036
4.45	1.28	14.15	6.95
-.005 (.005)	-.004 (.005)	-.013 (.005)*	-.012 (.005)*
.019 (.005)*	.016 (.005)*	.049 (.005)*	.038 (.005)*
.063	.067	.065	.065
.080	.049	.343	.238
16.17	9.60	96.94	58.24
-.005 (.005)	-.003 (.005)	-.012 (.006)*	-.011 (.005)*
.013 (.007)	.005 (.007)	.033 (.008)*	.021 (.008)*
.065	.069	.076	.073
.019	.002	.082	.039
3.67	0.40	16.65	7.48
-.004 (.004)	-.002 (.005)	-.010 (.005)	-.010 (.005)
.027 (.005)*	.028 (.006)*	.040 (.006)*	.032 (.006)*
.061	.065	.072	.070
.126	.122	.183	.132
26.76	25.78	41.69	28.23
-.004 (.004)	-.003 (.005)	-.011 (.005)*	-.010 (.005)*
.033 (.006)*	.031 (.007)*	.052 (.007)*	.041 (.007)*
.061	.065	.071	.069
.123	.098	.210	.149
26.18	2.28	49.41	32.49

Table 40.

Mixed Model Estimates: Correlations with Outcome Variables Regressed on Behavioral Characteristic Composite Scores.

	Deviance: Time 1	Deviance: Time 2	GPA
Agreeableness			
Intercept	.008 (.004)	.008 (.005)	-.005 (.003)
Slope	-.017 (.009)*	-.026 (.011)*	.023 (.005)*
SD: Slopes	.021	.036	.010
Model R^2	.099	.179	.080
Conscientiousness			
Intercept	.008 (.004)*	.008 (.004)*	-.005 (.003)
Slope	-.037 (.005)*	-.048 (.007)*	.019 (.004)*
SD: Slopes	.000	.026	.000
Model R^2	.265	.391	.047
Emotional Stability			
Intercept	.008 (.004)	.008 (.004)	-.006 (.003)
Slope	-.025 (.010)*	-.031 (.013)*	.024 (.005)*
SD: Slopes	.031	.044	.014
Model R^2	.150	.213	.084
Extraversion			
Intercept	.008 (.004)	.008 (.005)	-.005 (.003)
Slope	-.01 (.007)	-.001 (.007)	.011 (.004)*
SD: Slopes	.014	.000	.000
Model R^2	.065	.000	.014
Openness			
Intercept	.008 (.004)	.008 (.005)	-.005 (.003)
Slope	-.018 (.007)*	-.017 (.008)*	.014 (.006)*
SD: Slopes	.000	.000	.013
Model R^2	.032	.024	.040

Note. Standard error provided in parentheses. Estimates significant at $p < .05$ denoted by asterisk.

Table 40 (cont'd)

OCB: Time 1	OCB: Time 2	BARS Total: Time 1	BARS Total: Time 2
-.005 (.004)	-.003 (.005)	-.011 (.005)*	-.011 (.005)*
.031 (.010)*	.015 (.009)	.046 (.011)*	.032 (.011)*
.031	.012	.025	.031
.150	.034	.160	.145
-.005 (.005)	-.003 (.005)	-.011 (.004)*	-.011 (.005)*
.026 (.005)*	.024 (.007)*	.062 (.005)*	.049 (.006)*
.000	.020	.000	.000
.105	.161	.425	.291
-.005 (.004)	-.003 (.005)	-.011 (.005)*	-.011 (.005)*
.035 (.013)*	.014 (.013)	.057 (.013)*	.039 (.012)*
.050	.046	.034	.034
.238	.160	.197	.151
-.005 (.004)	-.003 (.005)	-.011 (.005)*	-.011 (.005)*
.036 (.006)*	.037 (.006)*	.054 (.008)*	.041 (.007)*
.000	.000	.026	.000
.168	.160	.315	.166
-.005 (.004)	-.003 (.004)	-.011 (.005)*	-.011 (.005)*
.050 (.009)*	.045 (.010)*	.071 (.008)*	.057 (.009)*
.029	.032	.000	.013
.291	.264	.277	.220

Table 41.

Model Estimates: *Criterion Correlations Regressed on Behavioral Characteristic Composite Scores*.

	Deviance: Time 1	Deviance: Time 2	GPA	OCB: Time 1	OCB: Time 2
Intercept	.008 (.003)*	.008 (.004)*	-.005 (.003)	-.005 (.004)	-.003 (.004)
Agreeableness	.025 (.013)	.001 (.014)	.017 (.008)*	.002 (.014)	.009 (.015)
Conscientiousness	-.038 (.007)*	-.057 (.008)*	.009 (.006)	.012 (.006)*	.008 (.007)
Emotional Stability	-.017 (.014)	.002 (.018)	.005 (.009)	.006 (.016)	-.016 (.018)
Extraversion	.002 (.008)	.026 (.009)*	.009 (.007)	.019 (.008)*	.024 (.010)*
Openness to Experience	-.008 (.011)	-.006 (.011)	-.012 (.009)	.024 (.012)*	.024 (.013)
SD: Agreeableness	.027	.015	.000	.000	.019
SD: Conscientiousness	.018	.022	.020	.000	.001
SD: Emotional Stability	.030	.061	.011	.036	.048
SD: Extraversion	.010	.018	.015	.005	.018
SD: Openness	.022	.017	.022	.021	.030
SD: Residual	.047	.048	.059	.053	.058
Model R^2	.453	.563	.186	.418	.366
AIC	-498.37	-474.00	-1100.46	-463.27	-431.13
BIC	-427.17	-402.80	-1010.85	-392.06	-359.93
-2LL	271.18	259.00	572.23	253.63	237.57
Number of Response Options	188	188	434	188	188
Number of Stems	36	36	80	36	36

Note. Standard error provided in parentheses. Estimates significant at $p < .05$ denoted by asterisk.

Table 41 (cont'd)

BARS Total: Time 1	BARS Total: Time 2
-.011* (.004)	-.011* (.004)
-.023 (.014)	-.025 (.016)
.046* (.007)	.038* (.008)
.027 (.015)	.017 (.015)
.028* (.009)	.018 (.010)
.020 (.012)	.024 (.013)
.023	.034
.019	.020
.030	.009
.025	.027
.033	.036
.050	.054
.661	.556
-474.05	-449.59
-402.85	-378.39
259.02	246.8
188	188
36	36

Table 42.

Model Estimates: *Criterion Correlations Regressed on Emotional Stability and Situational Characteristic Composite Scores*.

	Deviance: Time 1	Deviance: Time 2	OCB: Time 1
Intercept	.006 (.007)	.007 (.007)	-.004 (.007)
Emotional Stability	-.038 (.015)*	-.042 (.016)*	.054 (.022)*
Task Demands	-.001 (.012)	.000 (.012)	-.005 (.012)
Competition	.002 (.017)	.000 (.018)	-.001 (.018)
Interpersonal Relations	.002 (.018)	.007 (.020)	.001 (.019)
Morality, Integrity, Fairness	.005 (.013)	-.001 (.013)	-.002 (.013)
Individual-Emotional	.000 (.021)	.008 (.023)	.001 (.022)
Familiarity, Difficulty	.006 (.015)	.004 (.016)	-.003 (.016)
Social Pressure/Performance	-.003 (.020)	.006 (.022)	-.002 (.021)
Team Task Work	.000 (.008)	-.005 (.009)	.003 (.008)
Emotional Stability x Task Demands	-.029 (.028)	.012 (.029)	.017 (.040)
Emotional Stability x Competition	.007 (.040)	.001 (.042)	.025 (.060)
Emotional Stability x Interpersonal Relations	-.022 (.047)	-.133 (.049)*	-.086 (.068)
Emotional Stability x Morality, Integrity, Fairness	.030 (.029)	.083 (.030)*	.026 (.043)
Emotional Stability x Individual-Emotional	-.021 (.055)	-.031 (.056)	.013 (.080)
Emotional Stability x Familiarity, Difficulty	-.003 (.037)	.029 (.038)	-.027 (.054)
Emotional Stability x Social Pressure/Performance	.129 (.067)	.035 (.070)	.009 (.088)
Emotional Stability x Team Task Work	-.051 (.023)*	-.016 (.024)	.006 (.032)
SD: Emotional Stability	.022	.020	.060
SD: Residual	.058	.062	.061
AIC	-391.31	-369.11	-364.88
BIC	-326.58	-304.38	-300.15
-2LL	-547.50	-523.56	-513.11
R^2	.169	.207	.266
Number of Response Options	188	188	188
Number of Stems	36	36	36

Note. Standard error provided in parentheses. Estimates significant at $p < .05$ denoted by asterisk.

Table 42 (cont'd)

OCB: Time 2	BARS: Time 1	BARS: Time 2
.000 (.007)	-.009 (.008)	-.006 (.008)
.019 (.026)	.076 (.020)*	.056 (.020)*
-.006 (.013)	-.004 (.015)	-.009 (.014)
.006 (.019)	-.004 (.022)	.010 (.021)
.000 (.021)	-.002 (.024)	.007 (.023)
-.003 (.014)	-.001 (.016)	-.010 (.016)
.006 (.024)	-.006 (.027)	.000 (.027)
-.008 (.017)	-.009 (.019)	-.008 (.019)
.003 (.023)	.006 (.026)	.000 (.025)
.000 (.009)	.003 (.010)	.000 (.010)
.014 (.048)	.004 (.037)	.012 (.037)
-.018 (.072)	-.010 (.054)	-.003 (.053)
-.082 (.082)	.028 (.063)	-.031 (.062)
.038 (.051)	-.041 (.039)	-.003 (.038)
.014 (.095)	-.002 (.073)	-.011 (.072)
-.018 (.064)	-.033 (.049)	-.020 (.048)
-.014 (.103)	-.010 (.088)	-.001 (.086)
.010 (.038)	.018 (.031)	.022 (.030)
.078	.034	.035
.065	.075	.073
-334.52	-303.84	-314.38
-269.79	-239.11	-249.66
-478.08	-450.17	-461.54
.244	.210	.166
188	188	188
36	36	36

Table 43.

Mixed-Effects Means Model Estimates for Correlations with Individual Difference Characteristics and Criterion Outcomes.

	Estimate	*SE*	*SD*
Agreeableness	-.009*	.003	.000
Conscientiousness	-.005*	.003	.000
Emotional Stability	-.007*	.003	.000
Extraversion	-.002*	.003	.000
Openness	-.008*	.003	.000
Ability	-.014*	.003	.000
Experience: Business Courses	-.003*	.005	.000
Experience: Job Tenure	-.004*	.005	.000
Deviance: Time 1	.008*	.005	.000
Deviance: Time 2	.008*	.005	.000
GPA	-.005*	.004	.000
OCB: Time 1	-.005*	.005	.000
OCB: Time 2	-.003*	.005	.000
BARS Total: Time 1	-.011*	.005	.000
BARS Total: Time 2	-.011*	.005	.000
AIC	-12537.38		
BIC	-11525.59		
-2LL	-12988.81		

Note. *SE* = standard error, *SD* = standard deviation of mean estimates across item stems. Estimates significant at $p < .05$ denoted by asterisk.

Figure 1. Mean item ρ for situational characteristic ratings. (a) For all situational characteristic ratings (43). (b) For situational characteristics exhibiting non-zero ρ values (40).

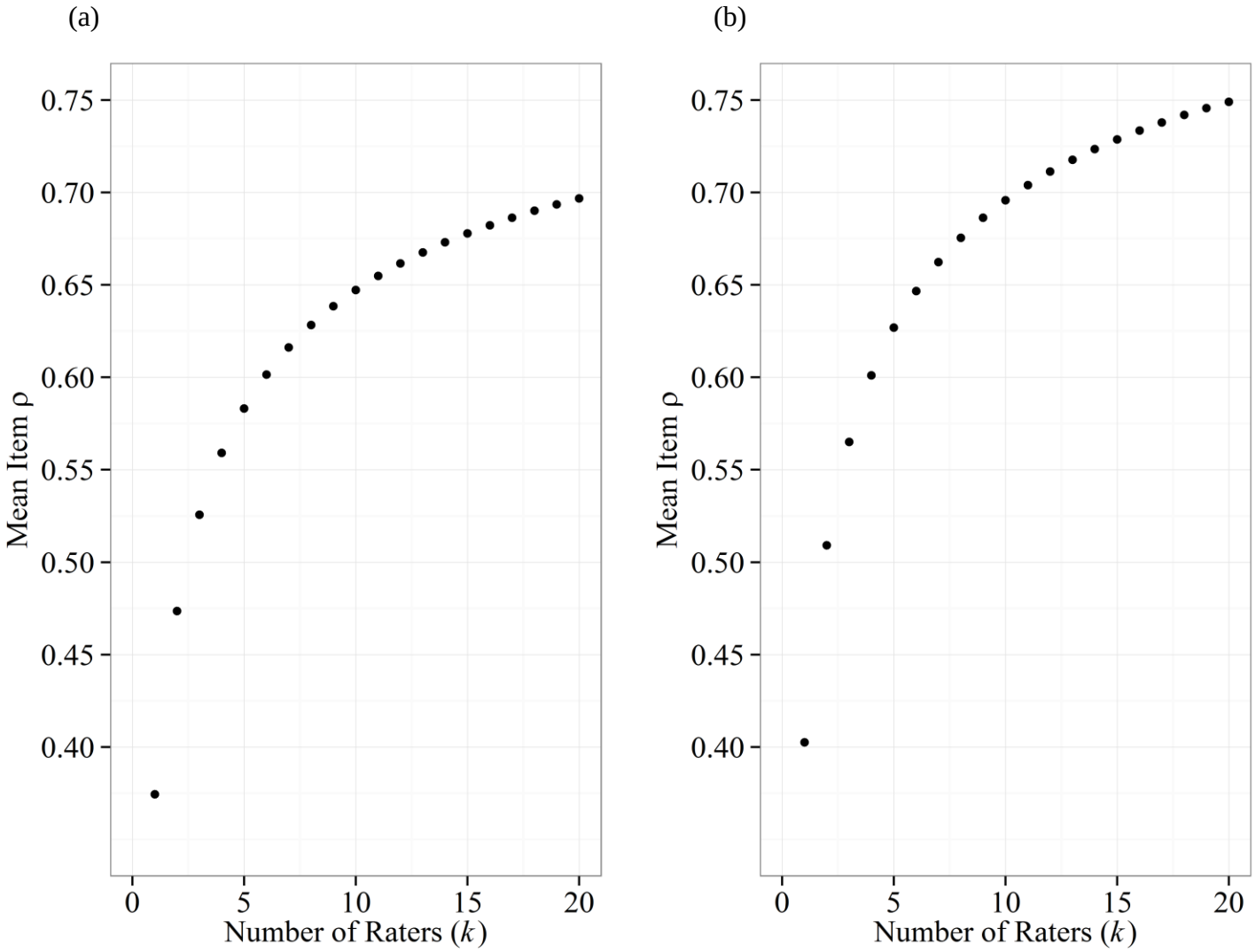


Figure 2. Mean item ρ for behavioral characteristic ratings.

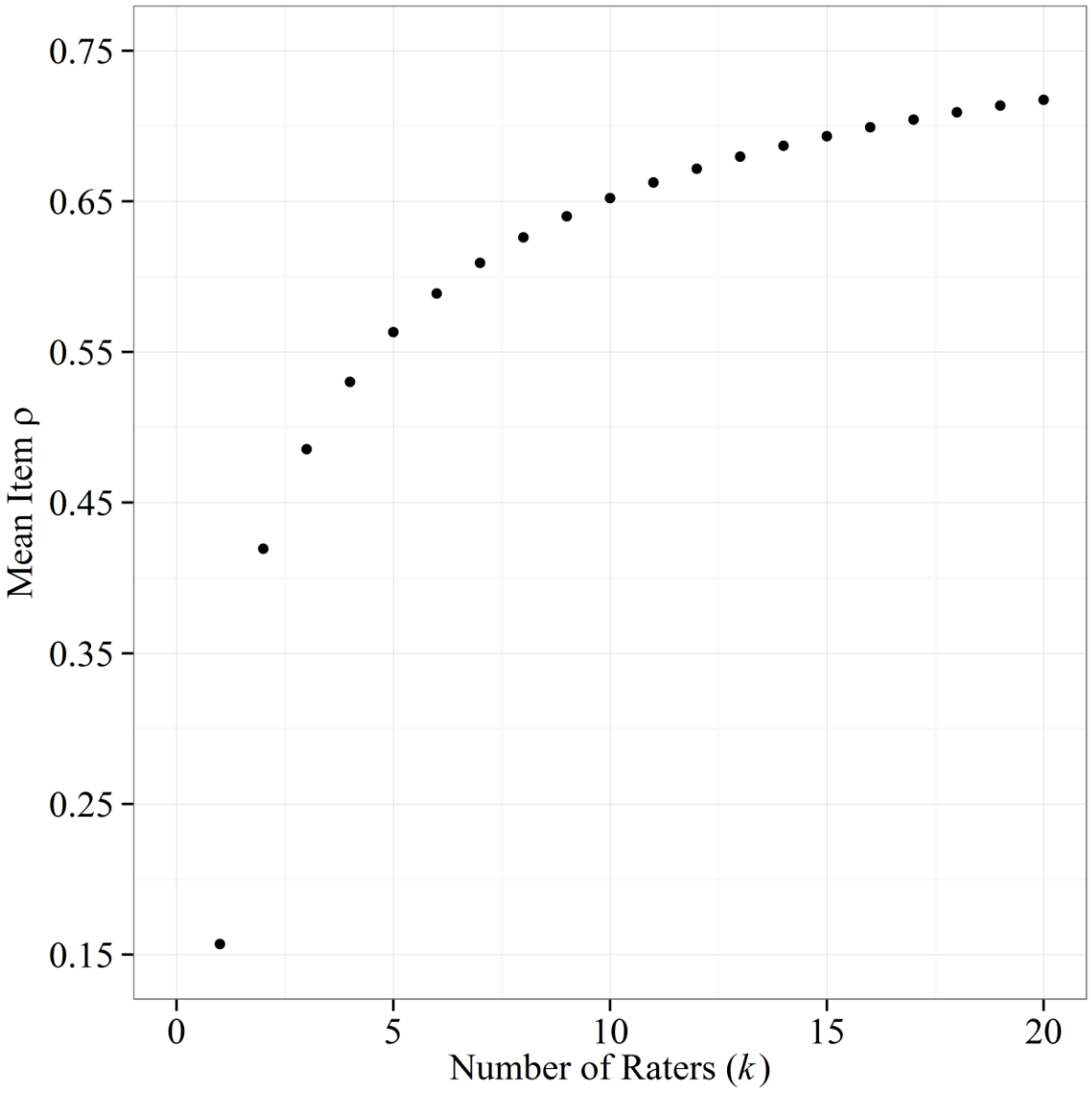


Figure 3. OLS-estimated slopes of FFM trait saturation correlations on FFM trait expression by item stem.

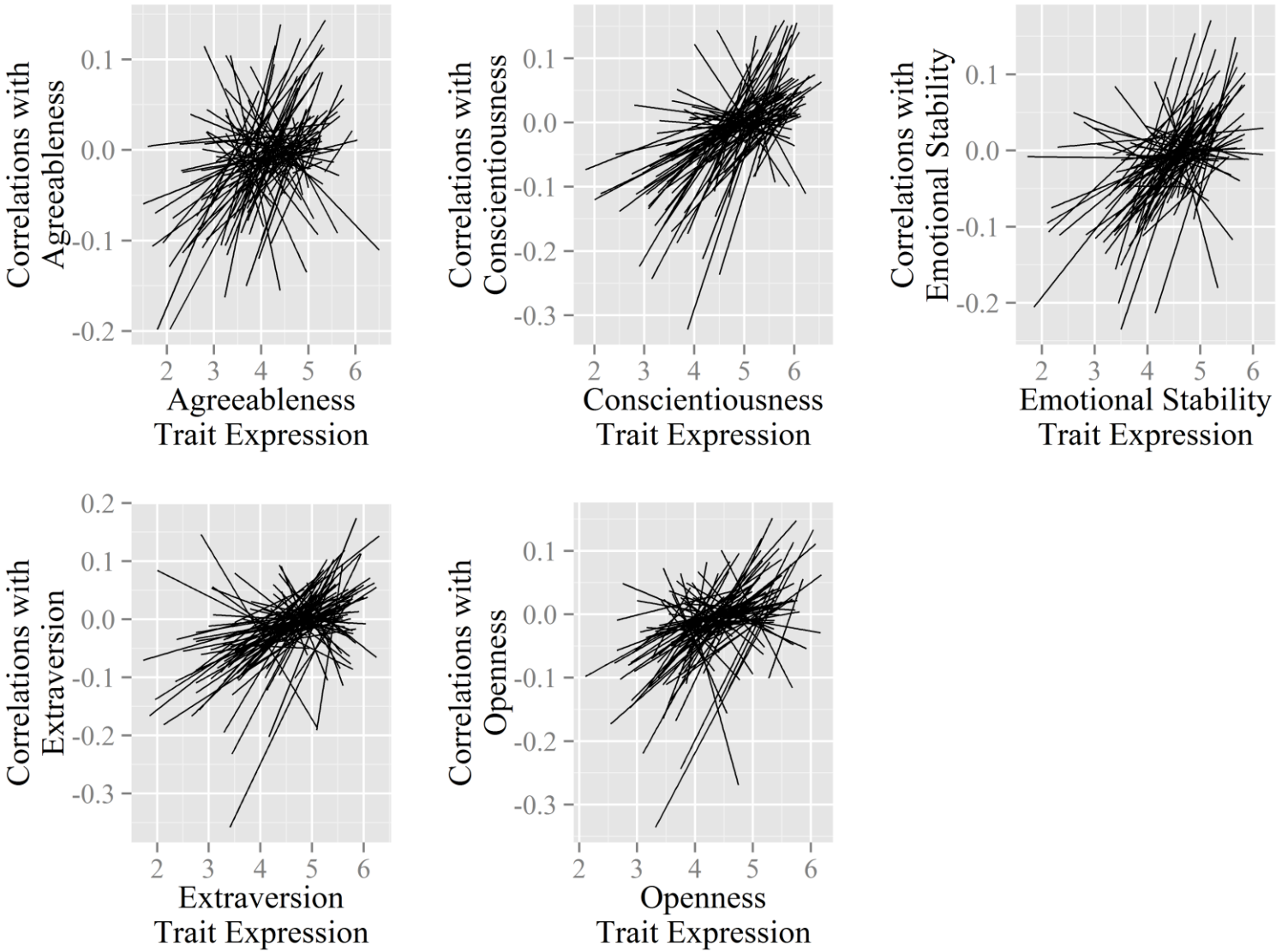


Figure 4. Density plot of FFM trait expression slopes.

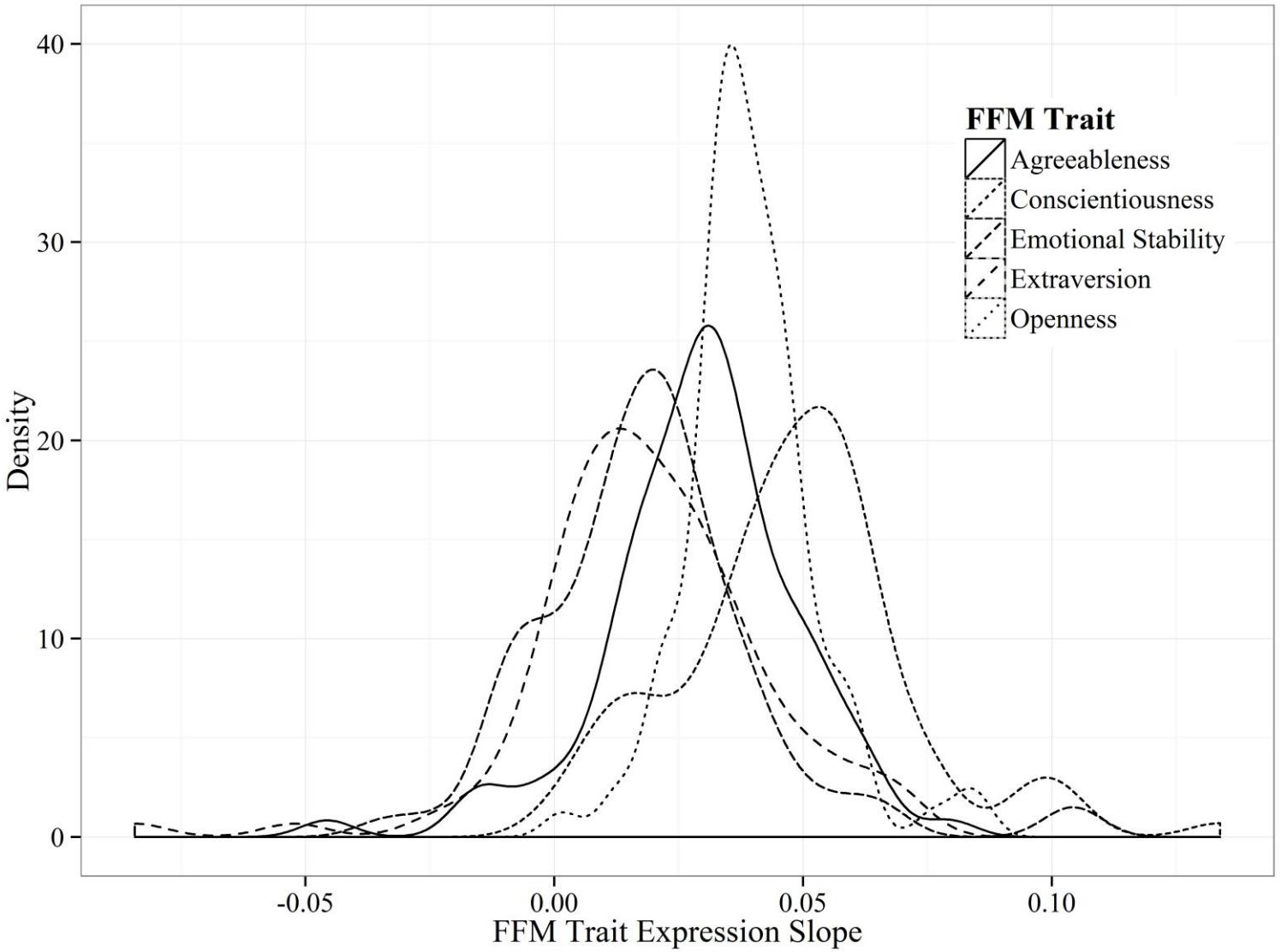


Figure 5. Simple slopes for FFM trait expression conscientiousness for predicting response option-level conscientiousness trait saturation.

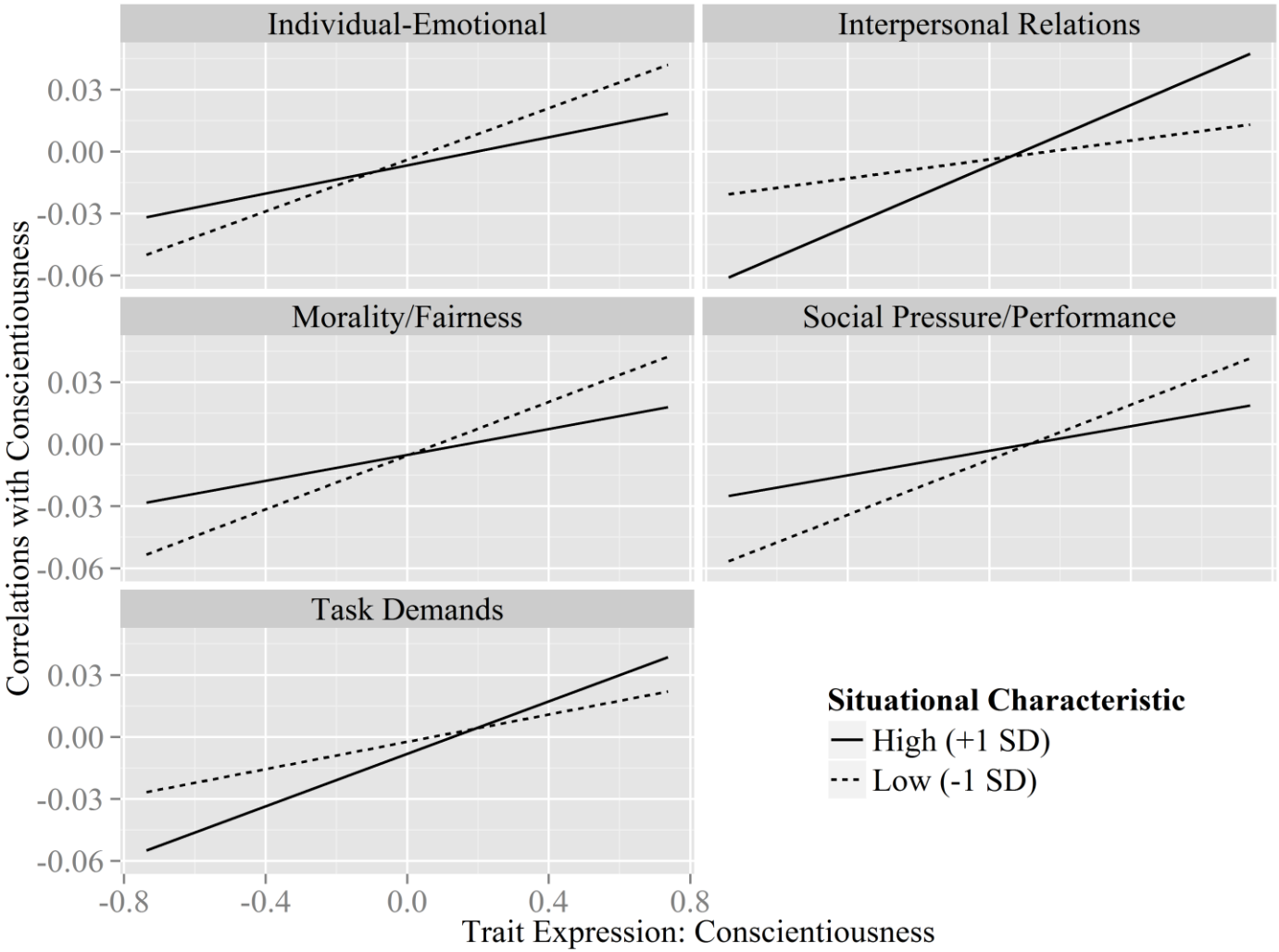


Figure 6. Simple slopes for FFM trait expression emotional stability for predicting response option-level emotional stability trait saturation.

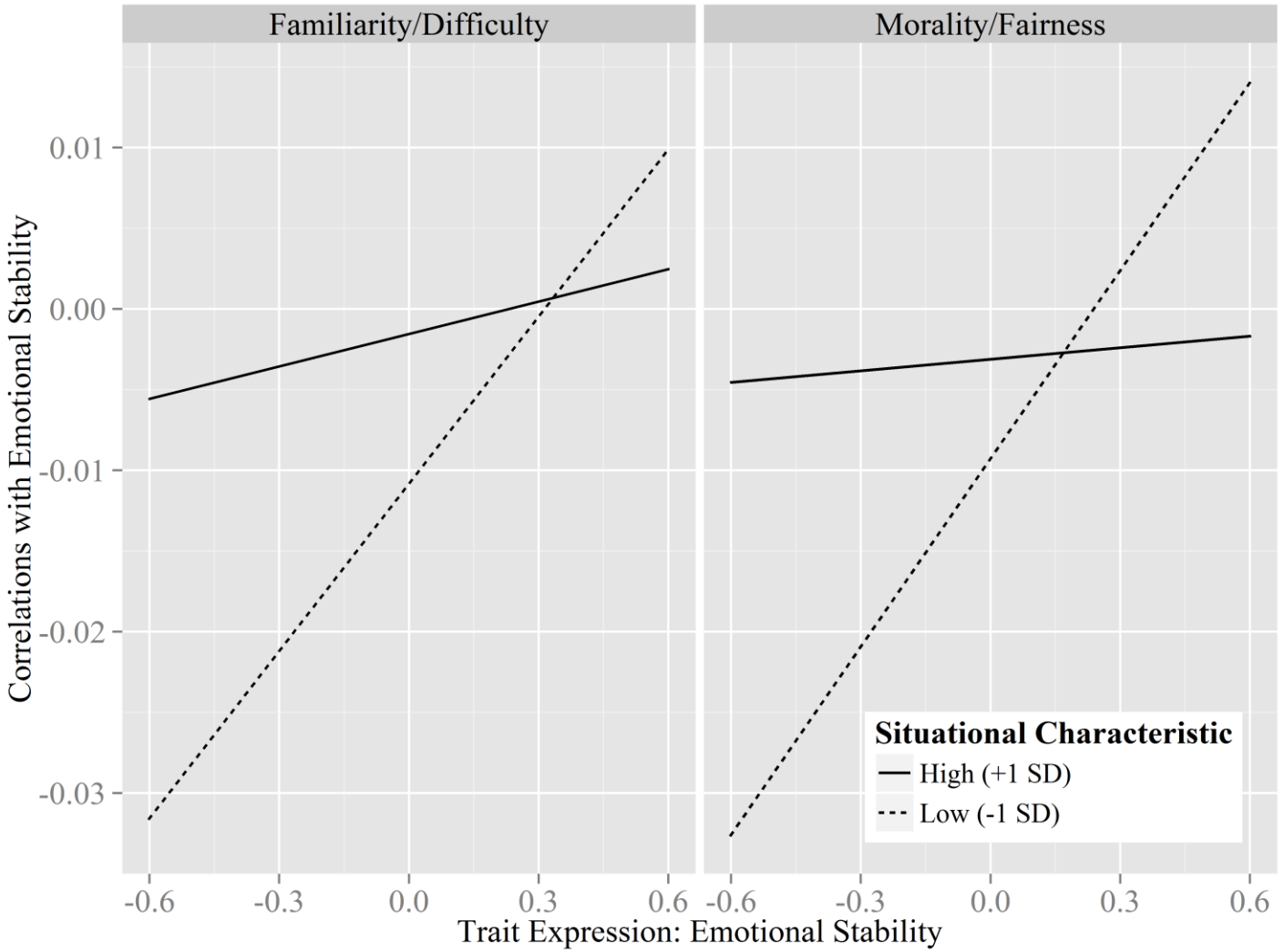


Figure 7. Simple slopes for FFM trait expression openness for predicting response option-level openness trait saturation.

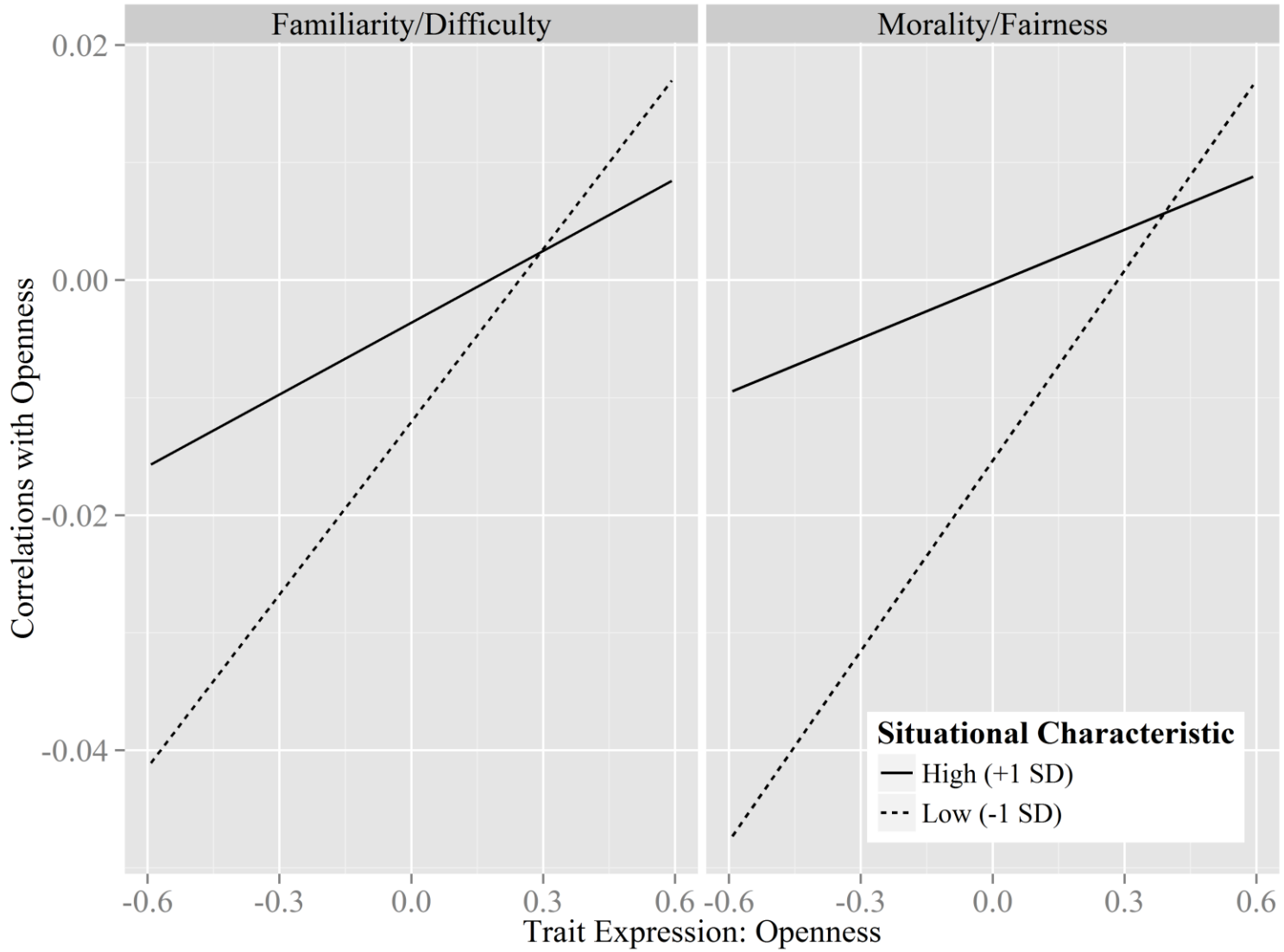


Figure 8. Density plot of emotional stability trait expression slopes.

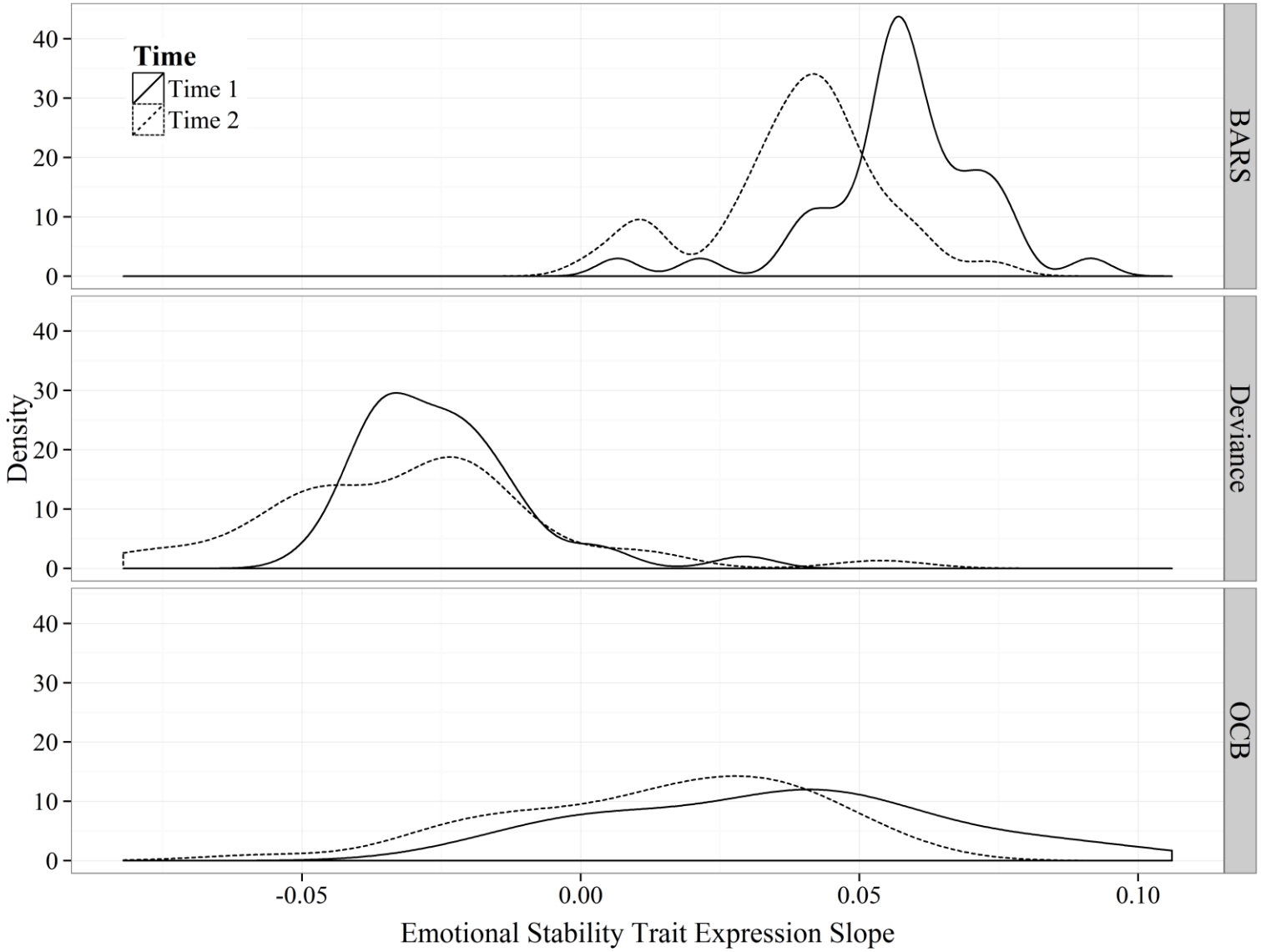


Figure 9. Simple slopes for FFM trait expression emotional stability for predicting criterion-related validities (*r*-to-*z*) between response option scores and Deviance: Time 1.

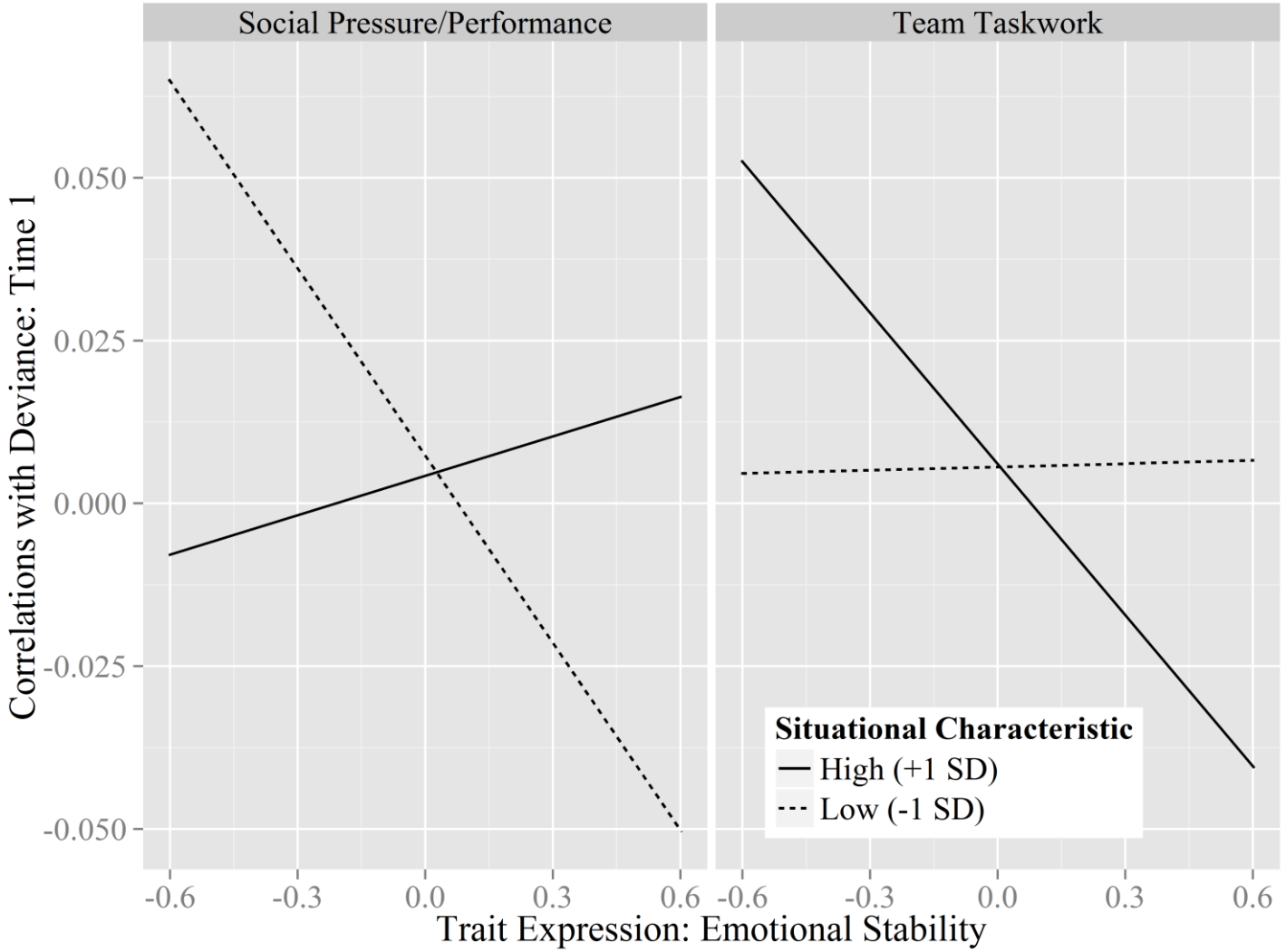
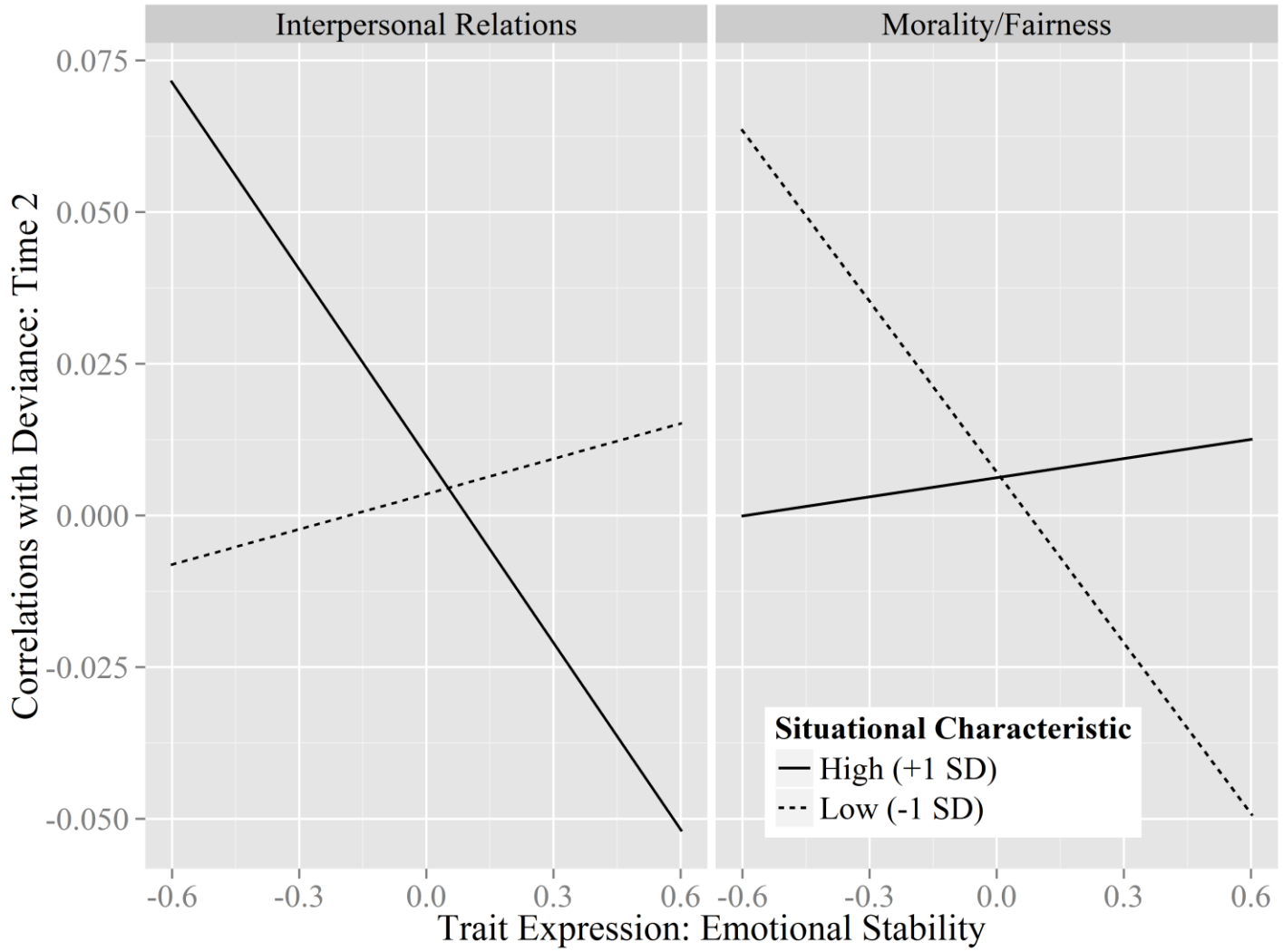


Figure 10. Simple slopes for FFM trait expression emotional stability for predicting criterion-related validities (*r*-to-*z*) between response option scores and Deviance: Time 2.



REFERENCES

REFERENCES

Anderson, L., & Wilson, S. (1997). The critical incident technique. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 89-112). Palo Alto, CA: Davies-Black Publishing.

Arthur, W. Jr., Day, E. A., McNelly, T. I., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56,* 125-154.

Arthur, W. Jr., Day, E. A., & Woehr, D. J. (2008). Mend it, don't end it: An alternate view of assessment center construct-related validity evidence. *Industrial and Organizational Psychology, 1,* 105-111.

Asher, J. J., & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27,* 519-533.

Bates, D., Maechler, M., & Bolker,B. (2011). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-42. http://CRAN.R-project.org/package=lme4.

Battisch, V. A., & Thompson, E. G. (1980). Students' perceptions of the college milieu: A multidimensional scaling analysis. *Personality and Social Psychology Bulletin, 6,* 74-82.

Beaty, J. C. Jr., Cleveland, J. N., & Murphy, K. R. (2001). The relation between personality and contextual performance in "strong" versus "weak situations. *Human Performance, 14,* 125-148.

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13,* 225-232.

Beckmann, N., Wood, R. E., & Minbashian, A. (2010). It depends how you look at it: On the relationship between neuroticism and conscientiousness at the within-and between-person levels of analysis. *Journal of Research in Personality, 44,* 593-601.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81,* 506-520.

Bem, D. J., & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. *Psychological Review, 85,* 485-501.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14,* 223-235.

Bledow, F., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62,* 229-258.

Block, J., & Block, J. H. (1981). Studying situational dimensions: A grand perspective and some limited empiricism. In D. Magnusson (Ed.)., *Toward a psychology of situations: An interactional perspective* (pp. 85-102). Hillsdale, NJ: Lawrence Erlbaum.

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6,* 1-21.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brown, R. D., & Hauenstein, N. M. A. (2005). Inter-rater agreement reconsidered: An alternative to r_{wg} indices. *Organizational Research Methods, 8,* 165-184.

Bycio, P., Alvares, K. M., & Hahn, J. (1987). Situational specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology, 72,* 463-474.

Cantor, N. (1981). Perceptions of situations: Situation prototypes and person-situation prototypes. In D. Magnusson (Ed.)., *Toward a psychology of situations: An interactional perspective* (pp. 229-244). Hillsdale, NJ: Lawrence Erlbaum.

Cervone, D., Shadel, W. G., & Jencius, S. (2001). Social-cognitive theory of personality assessment. *Personal and Social Psychology Review, 5,* 33-51.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143-159.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233-254.

Clause, C. S., Mullins, M. E., Nee, M. T., Pulakos, E., & Schmitt, N. (1998). Parallel test form development: A procedure for alternative predictors and an example. *Personnel Psychology, 51,* 193-208.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Harvey, V. S. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86,* 410-417.

Cooper, W. H., & Withey, M. J. (2009). The strong situation hypothesis. *Personality and Social Psychology Review, 13,* 62-72.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: PAR.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Fort Worth, TX: Harcourt Brace Jovanovich.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9,* 23-32.

De Meijer, L. A. L., Born, M. P., van Zielst, J., & van der Molen, H. T. (2010). Construct-driven development of a video-based situational judgment test for integrity. *European Psychologist, 15,* 229-236.

DeFruyt, F., & Mervielde, I. (1999). RIASEC types and big five traits as predictors of employment status and nature of employment. *Personnel Psychology, 52,* 701-727.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43,* 145-168.

Denison, D. R., Hart, S. L., & Kahn, J. A. (1996). From chimneys to cross-functional teams: Developing and validating a diagnostic model. *Academy of Management Journal, 39,* 1005-1023.

Eckes, T. (1995). Features of situations: A two-mode clustering study of situation prototypes. *Personality and Social Psychology Bulletin, 21,* 366-374.

Edwards, J. A., & Templeton, A. (2005). The structure of perceived qualities of situations. *European Journal of Social Psychology, 35,* 705-723.

Ekehammar, B. (1974). Interactionism in personality from a historical perspective. *Psychological Bulletin, 81,* 1026-1048.

Ekehammar, B., & Magnusson, D. (1973). A method to study stressful situations. *Journal of Personality and Social Psychology, 27,* 176-179.

Endler, N. S. (1966). Estimating variance components from mean squares for random and mixed effects analysis of variance models. *Perceptual and Motor Skills, 22,* 559-570.

Endler, N. S. (1981). Situational aspects of interactional psychology. In D. Magnusson (Ed.)., *Toward a psychology of situations: An interactional perspective* (pp. 361-373). Hillsdale, NJ: Lawrence Erlbaum.

Endler, N. S., & Hunt, J. M. (1966). Sources of behavioral variance as measured by the S-R inventory of anxiousness. *Psychological Bulletin, 65,* 336-346.

Endler, N. S., & Hunt, J. M. (1968). S-R inventories of hostility and comparisons of the proportions of variance from persons, responses, and situations for hostility and anxiousness. *Journal of Personality and Social Psychology, 9,* 309-315.

Endler, N. S., & Hunt, J. M. (1969). Generalizability of contributions from sources of variance in the S-R inventories of anxiousness. *Journal of Personality, 37,* 1-24.

Endler, N., Hunt, J. M., & Rosenstein, A. J. (1962). An S-R inventory of anxiousness. *Psychological Monographs, 76,* 1-33.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education, 15,* 49-74.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35,* 790-806.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360-392.

Epstein, S., & O'Brien, E. J. (1985). The person-situation debate in historical and current perspective. *Psychological Bulletin, 98,* 513-537.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality, 75,* 825-861.

Fleeson, W., & Noftle, E. E. (2008). Where does personality have its influence? A supermatrix of consistency concepts. *Journal of Personality, 76,* 1355-1385.

Forgas, J. P. (1976). The perception of social episodes: Categorical and dimensional representations in two different social milieus. *Journal of Personality and Social Psychology, 34,* 199-209.

Forgas, J. P. (1983). Episode cognition and personality: A multidimensional analysis. *Journal of Personality, 51,* 34-48.

Foster, J., Gaddis, H., & Hogan, J. (2012). Personality-based job analysis. In M. A. Wilson, W. Bennett Jr., S. G. Gibson, & G. M. Alliger (Eds.)., *The Handbook of Work Analysis: Methods, Systems, Applications and Science of Work Measurement in Organizations* (pp. 247-264). New York, NY: Routledge.

Fournier, M. A., Moskowitz, D. S., & Zuroff, D. C. (2008). Integrating dispositions, signatures, and the interpersonal domain. *Journal of Personality and Social Psychology, 94,* 531-545.

Frederiksen, N. (1972). Toward a taxonomy of situations. *American Psychologist, 27,* 114-123.

Frederiksen, N., Jensen, O., & Beaton, A. E. (1972). *Prediction of Organizational Behavior*. Elmsford, NY: Pergamon.

Fritzsche, B. A., Stagl, K. C., Salas, E., & Burke, C. S. (2006). Enhancing the design, delivery, and evaluation of scenario-based training: Can situational judgment tests contribute? In J.

A. Weekley and R. E. Ployhart (Eds.)., *Situational judgment tests: Theory, measurement and application* (pp. 301-318). Mahwah, NJ: Lawrence Erlbaum.

Funder, D. C. (2008). Persons, situations, and person-situation interactions. In L. Pervin, O. John, and R. Robins (Eds.), *Handbook of personality research* 3rd Ed. (pp. 568-582). New York: Guilford Press.

Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology, 60,* 773-794.

Furr, R. M., & Funder, D. C. (2004). Situational similarity and behavioral consistency: Subjective, objective, variable-centered, and person-centered approaches. *Journal of Research in Personality, 38,* 421-447.

Frederiksen, N. (1972). Toward a taxonomy of situations. *American Psychologist, 27,* 114-123.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. B. III., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493-511.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1,* 515-533.

Gelman, A., & Hill, J. (2009). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press: Cambridge.

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

Goldberg, L. R. (1999). *International personality item pool: A scientific collaboratory for the development of advanced measures of personality and other individual differences.* Retrieved from http://ipip.orgi.org/ipip/.

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

Goodenough, F. L. (1949). *Mental testing*. New York: Rinehart.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. Jr. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37,* 504-528.

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology, 55,* 137-163.

Hackman, J. R. (1987). The design of work teams. In J. W. Lorsch (Ed.), *Handbook of organizational behavior* (pp. 315-342). Englewood Cliffs, NJ: Prentice-Hall.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11,* 21-25.

Harman, R. P. (2012). Context analysis. In M. A. Wilson, W. Bennett Jr., S. G. Gibson, & G. M. Alliger (Eds.)., *The Handbook of Work Analysis: Methods, Systems, Applications and Science of Work Measurement in Organizations* (pp. 303-320). New York, NY: Routledge.

Hauenstein, N. M. A., Findlay, R. A., & McDonald, D. P. (2010). Using situational judgment tests to assess training effectiveness: Lessons learned evaluating military equal opportunity advisor trainees. *Military Psychology, 22,* 262-2881.

Heller, D., Perunovic, W. Q. E., & Reichman, D. (2009). The future of person-situation integration in the interface between traits and goals: A bottom-up framework. *Journal of Research in Personality, 43,* 171-178.

Highhouse, S., & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23,* 140-155.

Hill, G. J. (1989). An unwillingness to act: Behavioral appropriateness, situational constraint, and self-efficacy in shyness. *Journal of Personality, 57,* 871-890.

Holland, J. L. (1985). *Making vocational choices: A theory of vocational personalities and work environments* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement,*

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11.

Morgeson, F. P., & Humphrey, S. E. (2006). The work design questionnaire (WDQ): Developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology, 91,* 1321-1339.

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the behavioral indicators of conscientiousness (BIC). *Journal of Research in Personality, 44,* 501-511.

James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin, 81,* 1096-1112.

Jessor, R. (1956). Phenomenological personality theories and the data language of psychology. *Psychological Review, 63,* 173-180.

Kell, H. J., Rittmayer, A. D., Crook, A. E., & Motowidlo, S. J. (2010). Situational content moderates the association between the big five personality traits and behavioral effectiveness. *Human Performance, 23,* 213-228.

Kenrick, D. T., McCreath, H. E., Govern, J., King, R., & Bordin, J. (1990). Person-environment intersections: Everyday settings and common trait dimension. *Journal of Personality and Social Psychology, 58,* 685-690.

Klirs, E. G., & Revelle, W. (1986). Predicting variability from perceived situational similarity. *Journal of Research in Personality, 20,* 34-50.

Kobrin, J. L., Kim, Y., & Sackett, P. R. (2012). Modeling the predictive validity of SAT mathematics items using item characteristics. *Educational and Psychological Measurement, 72,* 99-119.

Kozlowski, S. W. J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77,* 161-167.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Elsevier, Inc.: Oxford.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and inter-rater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6,* 80-128.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about inter-rater reliability and inter-rater agreement. *Organizational Research Methods, 11,* 815-852.

Lewin, K. (1936). *Principles of topological psychology.* New York: McGraw-Hill.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90,* 442-452.

Lievens, F., Chasteen, C. S., Day, E. A., & Christiansen, N. D. (2006). Large-scale investigation of the role of trait activation theory for understanding assessment center convergent and discriminant validity. *Journal of Applied Psychology, 91,* 247-258.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86,* 1202-1222.

Lievens, F., & Sackett, P. R. (2007). Situational judgment tests in high-stakes settings: Issues and strategies with generating alternate forms. *Journal of Applied Psychology, 92,* 1043-1055.

Lindell, M. K., & Brandt, C. J. (1999). Assessing inter-rater agreement on the job relevance of a test: A comparison of the *CVI*, *T*, $r_{WG(J)}$, and $r^*_{WG(J)}$indexes. *Journal of Applied Psychology, 84,* 640-647.

Lord, C. G. (1982). Predicting behavioral consistency from an individual's perception of situational similarities. *Journal of Personality and Social Psychology, 42,* 1076-1088.

MacKenzie, W. I. Jr., Ployhart, R. E., Weekley, J. A., & Ehlers, C. (2010). Contextual effects on SJT responses: An examination of construct validity and mean differences across applicant and incumbent contexts. *Human Performance, 23,* 1-21.

Magnusson, D. (1971). An analysis of situational dimensions. *Perceptual and Motor Skills, 32,* 851-867.

Magnusson, D. (1981). Wanted: A psychology of situations. In D. Magnusson (Ed.)., *Toward a psychology of situations: An interactional perspective* (pp. 9-35). Hillsdale, NJ: Lawrence Erlbaum.

Magnusson, D., & Ekehammar, B. (1975). Perceptions of and reactions to stressful situations. *Journal of Personality and Social Psychology, 31,* 1147-1154.

Magnusson, D., & Ekehammar, B. (1978). Similar situations-similar behaviors? A study of the intraindividual congruence between situation perception and situation reactions. *Journal of Research in Personality, 12,* 41-48.

Marshall, M. A., & Brown, J. D. (2006). Trait aggressiveness and situational provocation: A test of the traits as situational sensitivities (TASS) model. *Personality and Social Psychology Bulletin, 32,* 1100-1573.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63-91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103-113.

McDaniel, M. A., Whetzel, D. L., Hartman, N. S., Nguyen, N. T., & Grubb, W. L. III. (2006). Situational judgment tests: Validity and an integrative model. In J. A. Weekley and R. E. Ployhart (Eds.)., *Situational judgment tests: Theory, measurement and application* (pp. 183-203). Mahway, NJ: Lawrence Erlbaum Associates.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30-46.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 13-23.

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

Meyer, R. D., Dalal, R. S., & Bonaccio, S. (2009). A meta-analytic investigation into the moderating effects of situational strength on the conscientiousness-performance relationship. *Journal of Organizational Behavior, 30,* 1077-1102.

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management, 36,* 121-140.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80,* 252-283.

Mischel, W. (1977). The interaction of person and situation. In D. Magnusson, & N. S. Endler (Eds.)., *Personality at the crossroads: Current issues in interactional psychology* (pp. 333-352). Hillsdale, NJ: Lawrence Erlbaum.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology, 55,* 1-22.

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.

Mischel, W., & Shoda, Y. (1998). Reconciling processing dynamics and personality dispositions. *Annual Review of Psychology, 49,* 229-258.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report RR-03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25,* 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *On the roles of task model variables in assessment design* (CSE Technical Report 500). Los Angeles, CA: Center for the Study of Evaluation.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond. R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15,* 363-389.

Monson, T. C., Hesley, J. W., & Chernick, L. (1982). Specifying when personality traits can and cannot predict behavior: An alternative to abandoning the attempt to predict single-act criteria. *Journal of Personality and Social Psychology, 43,* 385-399.

Moos, R. H. (1973). Conceptualizations of human environments. *American Psychologist, 28,* 652-665.

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

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640-647.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91,* 749-761.

Motowidlo, S. J., & Tippins, N. (1993). Further forms of the low-fidelity simulation in the form of a situational inventory. *Journal of Occupational and Organizational Psychology, 66,* 337-344.

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2006). Situational judgment in work teams: A team role typology. In J. A. Weekley and R. E. Ployhart (Eds.)., *Situational judgment tests: Theory, measurement and application* (pp. 319-343). Mahwah, NJ: Lawrence Erlbaum.

Mumford, T. V., Van Iddekinge, C. H., Morgeson, M. P., & Campion, M. A. (2008). The team role test: Development and validation of a team role knowledge situational judgment test. *Journal of Applied Psychology, 93,* 250-267.

Murray, H. (1938). *Explorations in personality*. New York: Oxford University Press.

Neidig, R. D., & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69,* 182-186.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13,* 250-260.

O'Connor, E. J., Peters, L. H., Pooyan, A., Weekley, J., Frank, B., & Erenkrantz, B. (1984). Situational constraint effects on performance, affective reactions, and turnover: A field replication and extension. *Journal of Applied Psychology, 69,* 663-672.

O'Reilly, C. A. III., Chatman, J., & Caldwell, D. F. (1991). People and organizational culture: A profile comparison approach to assessing person-organization fit. *Academy of Management Journal, 34,* 487-516.

Ostroff, C. (1993). Relationships between person-environment congruence and organizational effectiveness. *Group and Organization Management, 18,* 103-122.

Oswald, F. L., Friede, A. J., Schmitt, N., Kim, B. H., & Ramsay, L. J. (2005). Extending a practical method for developing alternate test forms using independent sets of items. *Organizational Research Methods, 8,* 149-164.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. R., & Gillespie, M. A. (2004). –Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187-207.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65,* 70-89.

Peters, L. H., Chassie, M. B., Lindholm, H. R., O'Connor, E. J., & Kline, C. R. (1982). The joint influence of situational constraints and goal setting on performance and affective outcomes. *Journal of Management, 8,* 7-20.

Peters, L. H., Fisher, C. D., & O'Connor, E. J. (1982). The moderating effect of situational control of performance variance on the relationship between individual differences and performance. *Personnel Psychology, 35,* 609-621.

Peters, L. H., & O'Connor, E. J. (1980). Situational constraints and work outcomes: The influences of a frequently overlooked construct. *Academy of Management Review, 5,* 391-397.

Peters, L. H., Fisher, C. D., & O'Connor, E. J. (1982). The moderating effect of situational control of performance variance on the relationship between individual differences and performance. *Personnel Psychology, 35,* 609-621.

Pervin, L. A. (1976). A free-response description approach to the analysis of person-situation interaction. *Journal of Personality and Social Psychology, 34,* 465-474.

Pervin, L. A. (1978). Definitions, measurements, and classifications of stimuli, situations, and environments. *Human Ecology, 6,* 71-105.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11,* 1-16.

Price, R. H. (1974). The taxonomic classification of behaviors and situations and the problem of behavior-environment congruence. *Human Relations, 27,* 567-585.

Price, R. H., & Bouffard, D. L. (1974). Behavioral appropriateness and situational constraint as dimensions of social behavior. *Journal of Personality and Social Psychology, 30,* 579-586.

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating inter-rater reliability. *Journal of Applied Psychology, 93,* 959-981.

R Development Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rapoport, A. (1978). *On the environment and the definition of the situation*. Paper presented at the Environmental Design Research Association 9 Conference, Tucson, AZ.

Rauthmann, J. F. (2011). Not only item content but also item format is important: Taxonomizing item format approaches. *Social Behavior and Personality, 39,* 119-128.

Rauthmann, J. F., & Denissen, J. J. A. (2011). I often do it vs. I like doing it: Comparing a frequency- and valency-approach to extraversion. *Personality and Individual Differences, 50,* 1283-1288.

Raymark, P. H., Schmit, M. J., & Guion, R. M. (1997). Identifying potentially useful personality constructs for employee selection. *Personnel Psychology, 50,* 723-736.

Reis, H. T. (2008). Reinvigorating the concept of situation in social psychology. *Personality and Social Psychology Review, 12,* 311-329.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349-359.

Rotter, J. B. (1954). *Social learning and clinical psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Sackett, P. R., & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401-410.

Sackett, P. R., & Dreher, G. F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology, 69,* 187-190.

Sackett, P. R., & Harris, M. M. (1988). A further examination of the constructs underlying assessment center ratings. *Journal of Business and Psychology, 3,* 214-229.

Saucier, G., Bel-Bahar, T., & Fernandez, C. (2007). What modifies the expression of personality tendencies? Defining basic domains of situation variables. *Journal of Personality, 75,* 479-503.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley and R. E. Ployhart (Eds.)., *Situational judgment tests: Theory, measurement and application* (pp. 135-155). Mahway, NJ: Lawrence Erlbaum Associates.

Schneider, B. (1975). Organizational climates: An essay. *Personnel Psychology, 28,* 447-479.

Schneider, B., & Reichers, A. E. (1983). On the etiology of climates. *Personnel Psychology, 36,* 19-39.

Schutte, N. S., Kenrick, D. T., & Sadalla, E. K. (1985). The search for predictable settings: Situational prototypes, constraint, and behavioral variation. *Journal of Personality and Social Psychology, 49,* 121-128.

Schutz, W. C. (1958). *FIRO: A three-dimensional theory of interpersonal behavior*. New York: Holt, Rinehart, & Winston.

Sherman, R. A., Nave, C. S., & Funder, D. C. (2010). Situational similarity and personality predict behavioral consistency. *Journal of Personality and Social Psychology, 99,* 330-343.

Shoda, Y. (2003). Individual differences in social psychology: Understanding situations to understand people, understanding people to understand situations. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.)., *The Sage Handbook of Methods in Social Psychology* (pp. 117-142). Thousand Oaks, CA: Sage.

Shoda, Y., Mischel, W., & Wright, J. C. (1993). The role of situational demands and cognitive competencies in behavior organization and personality coherence. *Journal of Personality and Social Psychology, 65,* 1023-1035.

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social Psychology, 67,* 674-687.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47,* 149-155.

Smith, R. E., Shoda, Y., Cumming, S. P., & Smoll, F. L. (2009). Behavioral signatures at the ballpark: Intraindividual consistency of adults' situation-behavior patterns and their interpersonal consequences. *Journal of Research in Personality, 43,* 187-195.

Snyder, M., & Ickes, W. (1985). Personality and social behavior. In G. Lindzey & E. Aronson (Eds.)., *Handbook of social psychology* (3rd ed.). (pp. 883-948). New York: Random House.

Stewart, G. L., & Nandkeolyar, A. K. (2007). Exploring how constraints created by other people influence intraindividual variation in objective performance measures. *Journal of Applied Psychology, 92,* 1149-1158.

Taylor, S. E., & Schneider, S. K. (1989). Coping and the simulation of events. *Social Cognition, 7,* 174-194.

Ten Berge, M. A., & De Raad, B. (1999). Taxonomies of situations from a trait perspective: A review. *European Journal of Personality, 13,* 337-360.

Ten Berge, M. A., & De Raad, B. (2001). The constructions of a job taxonomy of traits and situations. *European Journal of Personality, 15,* 253-276.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88,* 500-517.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34,* 397-423.

Thompson, C. I., Royce, S. E., & Bankart, C. P. (1987). Individuals' and groups' ratings of appropriateness of behavior. *Perceptual and Motor Skills, 64,* 255-260.

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependences. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 289-316). New York: Springer-Verlag.

Van Heck, G. L. (1984). The construction of a general taxonomy of situations. In H. Bonarius, G. L. Van Heck, and N. Smid. (Eds.)., *Personality psychology in Europe: Theoretical and empirical developments* (pp. 149-164). Lisse, Netherlands: Swets & Zeitlinger.

Van Heck, G. L. (1989). Situation concepts: Definitions and classification. In P. J. Hettema (Ed.)., *Personality and environment: Assessment of human adaptation* (pp. 53-69) and pp. 241-259). Chichester: Wiley.

Van Heck, G. L., Perugini, M., Caprara, G.-V., & Froger, J. (1994). The big five as tendencies in situations. *Personality and Individual Differences, 16,* 715-731.

Villanova, P., & Roman, M. A. (1993). A meta-analytic review of situational constraints and work-related outcomes: Alternative approaches to conceptualization. *Human Resource Management Review, 3,* 147-175.

Wagerman, S. A., & Funder, D. C. (2008). Personality psychology of situations. In P. J. Corr and G. Matthews (Eds.), *Cambridge handbook of personality* (pp. 27-42). Cambridge: Cambridge University Press.

Wageman, R., Hackman, J. R., & Lehman, E. (2005). Team diagnostic survey: Development of an instrument. *The Journal of Applied Behavioral Science, 41,* 373-398.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15,* 22-29.

Wallace, J. (1966). An abilities conception of personality: Some implications for personality measurement. *American Psychologist, 21,* 132-138.

Wang X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and application. *Applied Psychological Measurement, 26,* 109-128.

Wang, W.-C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29,* 296-318).

Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the positive and negative affect schedule – expanded form.* Iowa City: The University of Iowa.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063-1070.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50,* 25-49.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52,* 679-700.

Weekley, J. A., & Ployhart, R. E. (2005). Situational judgment: Antecedents and relationships with performance. *Human Performance, 18,* 81-104.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley and R. E. Ployhart (Eds.)., *Situational judgment tests: Theory, measurement and application* (pp. 157-182). Mahway, NJ: Lawrence Erlbaum Associates.

Werner, P. D., & Pervin, L. A. (1986). The content of personality inventory items. *Journal of Personality and Social Psychology, 51,* 622-628.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52,* 372-376.

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology, 75,* 315-321.

Wish, M., Deutsch, M., & Kaplan, S. J. (1976). Perceived dimensions of interpersonal relations. *Journal of Personality and Social Psychology, 33,* 409-420.

Wonderlic. (1999). *Wonderlic personnel test and scholastic level exam user's manual.* Libertyville, IL: Author.

Wong, S. P., & McGraw, K. O. (1999). Confidence intervals and *F* tests for intraclass correlations based on three-way random effects models. *Educational and Psychological Measurement, 59,* 270-288.

Wright, J. C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology, 53,* 1159-1177.

Yang, Y., Read, S. J., & Miller, L. C. (2006). A taxonomy of situations from Chinese idioms. *Journal of Research in Personality, 40,* 750-778.

Zayas, V., & Shoda, Y. (2009). Three decades after the personality paradox: Understanding situations. *Journal of Research in Personality, 43,* 280-281.

Zhou, H., Muellerleile, P., Ingram, D., & Wong, S. P. (2011). Confidence intervals and *F* tests for intraclass correlation coefficients based on three-way mixed effects models. *Journal of Educational and Behavioral Statistics,* 36, 638-671.

Zickar, M. J., & Ury, K. L. (2002). Developing an interpretation of item parameters for personality items: Content correlates of parameter estimates. *Educational and Psychological Measurement, 62,* 19-31.