

**EXAMINATION OF THE EFFECTS OF WOULD-DO AND SHOULD-DO
INSTRUCTION SETS ON THE CONSTRUCT-RELATED VALIDITY OF
SITUATIONAL JUDGMENT TESTS**

A Dissertation

by

JUAN CARLOS BATARSE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee, Winfred Arthur, Jr.
Committee Members, Stephanie C. Payne
Deidra J. Schleicher
Steven M. Smith
Head of Department, Heather C. Lench

May 2018

Major Subject: Industrial/Organizational Psychology

Copyright 2018 Juan Carlos Batarse

ABSTRACT

It is purported that taking a should-do situational judgment test (SJT) engenders a more cognitive/knowledge-based orientation; in contrast to a would-do SJT which engenders a more personality/behavioral-based orientation. Although this effect is widely regarded as received doctrine, a detailed review of the literature indicates there have not been any direct tests of this effect using a *single* word manipulation with a manipulation check. Therefore, despite previous research on the influence of response instruction sets on the construct-related validity of SJTs, further examination of this issue is warranted.

Consequently, this study examines the SJT "would-do"/"should-do" response instruction effect. The present study accomplishes a number of objectives. First, it examines the recall of the focal words and the effectiveness of an intervention (i.e., increasing the saliency of the focal word) to increase recall, and by inference, processing. Second, it seeks to replicate the posited response instruction effect using this single word manipulation with a manipulation check. Third, it further extends the literature by examining how the construct assessed influences the instruction set effect.

Results indicated that participants failed to accurately recall the focal word (i.e., "would" or "should") in the response instructions, even with a modified stronger degree of saliency (i.e., **bold/underline**). Furthermore, the pattern of results currently purported in the literature regarding the effect of response instructions on the construct-related validity of SJTs was not replicated with a single word manipulation. Finally, the construct assessed did not influence the instruction set effect or lack thereof.

However, the findings suggest that asking participants to recall the focal word after the first SJT item provides a priming effect and thus, can serve as an intervention to increase the likelihood that test-takers process the focal word. Furthermore, given the present study's failure to replicate the instruction set effect with a single word manipulation, the importance that can be placed on the instruction set's influence on the construct-related validity of SJTs is limited. Finally, a preliminary trend of a repetition effect was found such that the more SJT items a participant was exposed to before responding to a manipulation check, the higher the recall rate for that particular manipulation check.

DEDICATION

I dedicate this dissertation to my parents in appreciation for their unconditional love and support.

ACKNOWLEDGEMENTS

There are many individuals whose contributions were important to the completion of this work. First, I recognize my advisor Dr. Winfred Arthur, Jr. for his significant support and guidance on this project and throughout my graduate career. His mentoring has taught me (among other things) the value of (1) using systematic methods in science, and (2) precise communication. Second, thank you to all of my other committee members, Dr. Stephanie Payne, Dr. Steve Smith, and Dr. Deidra Schleicher, for providing guidance and thoughtful criticism on this project. Third, I acknowledge the help of the Industrial/Organizational doctoral program at Texas A&M University, whose members (both faculty and students) showed support by providing useful feedback on various occasions. Fourth, I specifically thank Zachary Traylor for providing insightful thoughts and feedback on diverse topics related to this project. Finally, I'd like to acknowledge Christen Dovalina, Craig White, and Inchul Cho for their help in developing and refining the cognitive SJT used in the present study.

CONTRIBUTORS AND FUNDING SOURCES

This work was supervised by a dissertation committee consisting of Professor Stephanie C. Payne, Professor Steven M. Smith and Professor Winfred Arthur, Jr. of the Department of Psychology, and Professor Deidra J. Schleicher of the Department of Management in the Mays Business School.

All work for the dissertation was completed by the student, under the advisement of Winfred Arthur, Jr. of the Department of Psychology.

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

	Page
ABSTRACT	ii
DEDICATION	iv
ACKNOWLEDGEMENTS	v
CONTRIBUTORS AND FUNDING SOURCES.....	vi
LIST OF FIGURES.....	x
LIST OF TABLES	xi
INTRODUCTION AND LITERATURE REVIEW.....	1
Strengths of Situational Judgment Tests	2
Situational Judgment Test-Use in Practice.....	4
Methods vs. Constructs	6
Research on Situational Judgment Test Design Characteristics and Sample-Type Effects.....	7
Interaction modality.	8
Fidelity.	9
Stem complexity.....	9
Mode of presentation.....	10
Response format.	11
Scoring key development.	12
Scoring strategy.....	13
Sample-type effects.	15
Response instructions.....	15
Research on Situational Judgment Test Response Instruction Sets	16
Faking.....	18
Construct-related validity.....	20

Present Study.....	24
Construct-Related Validity: Would vs. Should.....	27
False memory.....	27
Anticipation.....	28
Eye movements during word processing.....	30
Construct-Related Validity: Constructs Assessed and Instruction Sets	35
METHOD.....	40
Participants	40
Measures.....	40
General mental ability (GMA).....	40
Agreeableness.....	41
Cognitive SJT.....	41
Agreeableness (noncognitive) SJT.....	45
Manipulation check item.....	45
Demographics.....	47
Design and Procedure.....	47
RESULTS.....	50
Hypothesis Testing.....	50
DISCUSSION AND SUMMARY	67
Implications	73
Implications for science.....	73
Implications for practice.....	74
Limitations, Directions for Future Research, and Conclusions	76
Limitations.....	76
Directions for future research.....	78
Conclusion.....	80
REFERENCES	81

APPENDIX A	90
APPENDIX B	92

LIST OF FIGURES

	Page
<i>Figure 1.</i> Variety of operationalizations of "would-do" and "should-do" instruction sets used in the literature.	18
<i>Figure 2..</i> Matrix of 8 study conditions.	32
<i>Figure 3.</i> Sequence of SJT items and manipulation checks presented to participants..	47
<i>Figure 4.</i> Manipulation check pass rates by condition.....	54
<i>Figure 5.</i> Manipulation check pass rates by location and construct assessed.....	55

LIST OF TABLES

	Page
Table 1. Descriptive Statistics and Intercorrelations for All Study Variables	51
Table 2. Manipulation Check Pass Rates and Associated Chi Square Tests	52
Table 3. Manipulation Check Pass Rates and Associated Chi Square Tests	57
Table 4. Pass Rate Comparisons between Standard Saliency and Bold/Underline Conditions	58
Table 5. Pass Rate Comparisons between Standard Saliency and Bold/Underline Conditions by Construct.....	59
Table 6. Hypothesis 3: Correlations between Would-do and Should-do SJT scores and GMA and Agreeableness.....	60
Table 7. Hypothesis 3: Correlations between Would-do and Should-do SJT scores and GMA and Agreeableness (excluding participants who failed post-5 th item MC)	62
Table 8. Research Question 1: Correlations between Would-do and Should-do SJT scores and GMA and Agreeableness (excluding participants who failed post-5 th item MC) by Construct.....	65
Table 9. Research Question 1: Correlations between Would-do and Should-do SJT scores and GMA and Agreeableness by Construct	66
Table B1. Hypotheses 4 and 5: Correlations between GMA and Agreeableness and SJT Scores with Theoretically-aligned Combination of Instruction Set and Construct Assessed.....	93

INTRODUCTION AND LITERATURE REVIEW

This study examines the effects of would-do and should-do response instructions on the construct-related validity of situational judgment tests (SJTs). The SJT is a selection method that typically presents applicants with situations (e.g., work-related situations) and response options and asks them how they would or should respond to the situation (see Appendix A for a sample item). Responses are then scored with a predetermined key.

The use of SJTs is increasingly common in the workplace today, especially in personnel selection and other organizational decision-making contexts. As a predictor method, the SJT is characterized by a number of design characteristics (e.g., interaction modality, fidelity, stem complexity, mode of presentation, response instructions, response format, scoring key development, scoring strategy, and sample-type effects). One SJT design characteristic is the response instruction—specifically, how test-takers are instructed to respond to the items—that is, how they would (would-do response instruction sets) or should (should-do response instruction sets) respond to the situation presented in the scenario. This particular design characteristic is the focus of the present study because it is advanced here that the received doctrine that serves as the basis for its posited effects has not been unambiguously examined and thus supported in the literature. Specifically, the research on the extent to which response instructions influence the construct-related validity of SJTs has several methodological problems that call into question the validity of subsequent results and conclusions. Furthermore, any influence that the construct measured might have on this effect has not been examined.

Consequently, the present study examines how the use of would-do versus should-do response instructions influences the construct-related validity of SJTs.

Concerning the organization of this dissertation, first, the strengths of SJTs are reviewed. Second, a number of overall issues germane to their usage (e.g., typical setting) as well as a brief history of SJTs are presented. Third, given its centrality to the issues at hand, the fundamental distinction between predictor methods and predictor constructs is highlighted. Fourth, research on SJT design characteristics and sample-type effects is reviewed. Fifth, a specific and detailed review of the focal design characteristic—response instructions—is provided. Sixth, the study’s hypotheses are developed and presented. Finally, after the research design and method, the results are presented and discussed, followed by the potential implications for science and practice.

Strengths of Situational Judgment Tests

The interest in SJTs as predictors of work performance is likely due to a number of advantages that they have (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Before discussing these strengths, it should be noted that some of the studies reviewed below are conceptually problematic because they can be characterized as having the classic construct/method confound, which makes it difficult to conclude whether the effects found are construct- or method-effects (Arthur & Villado, 2008). Consonant with this, SJTs are measurement methods and are discussed as such in the present study; although, some of the summaries presented below are characterized by the construct/method confound.

As noted above, the popularity of SJTs can be attributed to a number of documented advantages. First, they have moderately strong criterion-related validity ($\rho = .34$; McDaniel et al., 2001). However, McDaniel, Hartman, Whetzel, and Grubb's (2007) meta-analysis, after updating the primary studies by including several new unpublished studies, reported a somewhat lower validity ($\rho = .26$). Christian, Edwards, and Bradley (2010) advanced the SJT validity literature by examining the validity of SJTs at the construct level. When the constructs assessed were accounted for, SJTs still displayed moderate to moderately strong criterion-related validity for a variety of constructs: job knowledge and skills ($M_\rho = .19$; operational validity corrected for criterion unreliability), interpersonal skills ($M_\rho = .25$), teamwork skills ($M_\rho = .38$), leadership ($M_\rho = .28$), personality composites (including various composites of conscientiousness, agreeableness, emotional stability, adaptability, and integrity; $M_\rho = .43$), and conscientiousness ($M_\rho = .24$; Christian et al., 2010). These findings indicate that the SJT offers useful levels of criterion-related validity.

Second, SJTs demonstrate greater face validity, which has been linked to increased test-taking motivation (Bauer & Truxillo, 2006). Third, SJTs tend to exhibit smaller racial and sex subgroup differences. The mean difference between African-Americans and Whites on SJTs tends to be between .20 and 1.20 favoring Whites (Nguyen & McDaniel, 2003). Whetzel, McDaniel, and Nguyen's (2008) meta-analysis showed an African-American/White standardized mean difference (d) of .38, and even lower mean differences for Hispanic/White and Asian/White comparisons, .24 and .29, respectively. Given that the widely-cited, standardized mean difference in general

mental ability (GMA) test scores between African-Americans and Whites is about 1.00 (Roth, Bevier, Bobko, Switzer, & Tyler, 2001), these *ds* are considerably smaller. In addition, given that the Asian/White standardized mean difference in GMA is -.20 (in favor of Asians), it is notable that there is a difference in the opposite direction on SJTs (Ployhart & Holtz, 2008). Finally, the Hispanic/White standardized mean difference on SJTs (.24) is lower than the Hispanic/White mean difference on GMA, which ranges from .58 to .83.

Men and women have traditionally been known to differ on certain individual differences. For example, women tend to be more agreeable ($r = .17$), more open ($r = .13$), and less emotionally stable ($r = -.21$; Willie, Fruyt, & Feys, 2010). However, male/female differences on cognitive ability are negligible (Ployhart & Holtz, 2008). Regarding SJTs, few studies report sex differences, and if they are reported, then they are typically in favor of women. For example, Weekley and Jones (1999) reported a mean difference of .30 in favor of women, and Whetzel et al.'s (2008) meta-analysis also reported a standardized mean difference of .11 in favor of women. As demonstrated by these results, it is clear that SJTs can have value in selection contexts in terms of criterion-related validity, and potentially, reduced subgroup differences.

Situational Judgment Test-Use in Practice

SJTs can be traced back to the early 1920's, with one of the earliest examples being the George Washington Social Intelligence Test, which gave respondents multiple solutions to a presented scenario (Ployhart & MacKenzie, 2011). Its use continued through World War II, when soldiers were given SJTs to assess their judgment. These

tests were used for both military and civil service examinations. Throughout the mid-1900's, SJTs were used to predict managerial potential of employees in corporations. In the late 1900's, SJTs were reintroduced as a low fidelity simulation by Motowidlo, Dunnette, and Carter (1990), which has since generated a lot of research on the topic.

SJTs are used both in academia and applied/industry settings. This includes the public sector, private sector, and the military. SJTs are designed by both consulting firms and industrial/organizational (I/O) researchers and practitioners, and are used in both selection and placement, and training and development contexts (Ployhart, 2006). The general structure of SJTs is quite standardized, and with some variation in detail, generally takes the following form—a hypothetical scenario (i.e., work-related or otherwise) is presented, and the respondent is asked to pick or evaluate the response options to the scenario (McDaniel et al., 2001). Again, see Appendix A for an illustrative example.

A number of other issues are germane to how SJTs are used in practice. First, SJTs are typically context-bound, although the situations or scenarios can be described fairly generically to apply to different contexts (Ployhart & MacKenzie, 2011). That being said, Motowidlo, Ghosh, Mendoza, Buchanan, and Lerma (2016) developed a generic, *context-independent* SJT of prosocial implicit trait policy that can predict behavior in situations that are different than the contexts of the SJT items. Arthur (2017a) makes similar arguments for context-independent, construct-laden SJTs as well. Second, items can present response options that contain multiple and competing goals, which must be balanced by the respondent. Several plausible options may exist, but

there is always a *best* response (whether it be a single-response, multi-response, or a particular rank ordering). Being characterized by a particular level of the measured construct tends to be required for the test-taker to be insightful about which is the best response. Third, test developers must address issues related to item development, which includes situation generation, response option and key generation, and scoring (Ployhart & MacKenzie, 2011). Like any standardized assessment or measurement tool, generating appropriate item stems and appropriate response options are both critical to the valid measurement of the intended construct.

Methods vs. Constructs

It is important to distinguish between predictor constructs and predictor methods as it allows for isolation of variance due to predictor constructs and methods, which facilitates meaningful comparisons (Arthur & Villado, 2008). Predictor constructs are the domain that is being sampled (i.e., content of *what* is being measured), whereas predictor methods are the specific techniques by which this domain-relevant information is measured (i.e., *how* assessments measure what they are designed to measure). That is, predictor constructs are the behavioral domain that is being sampled (e.g., GMA, agreeableness, and adaptability), whereas predictor methods are the techniques or methods used to obtain information regarding the behavioral domain (e.g., interviews, paper-and-pencil tests, and assessment centers).

Because SJTs are methods, they can be designed to measure a variety of different constructs (e.g., cognitive ability, agreeableness). SJTs that measure different constructs will by definition have different construct-related validity. Therefore, it is erroneous to

talk about the construct-related validity of SJTs in general; rather, one must talk about the construct-related validity of an SJT given the construct it measures. Consequently, because they are predictor methods, an examination of the psychometric effects of varying response instruction sets requires considering the construct being measured.

Christian et al.'s (2010) meta-analysis of the criterion-related validity of SJTs at the construct-level provided a summary of the constructs that SJTs have been used to measure in the extant literature. As Christian et al. discuss, this construct-based approach carries numerous advantages. For example, focusing on the construct instead of the method allows for an examination of *why* that measure or construct predicts work performance. Next, and related to the previous advantage, a reporting of constructs allows for the development of SJTs that can be predict performance across situations. This is difficult to do when data are at the method level, especially given that some predictor constructs are more relevant to some jobs than others. Finally, a construct-based approach allows researchers to make more precise comparisons between different selection methods, especially when measuring the same constructs. That being said, Christian et al. reported that one-third of the studies they reviewed did not report the constructs measured by the SJT. Therefore, there is a need for a focus on the constructs assessed that is currently not being addressed in the literature.

Research on Situational Judgment Test Design Characteristics and Sample-Type Effects

In selection contexts, simulations are a way to examine how knowledge, skills, and abilities (KSAs) translate into performance (Lievens & Patterson, 2011). Candidates

typically perform a set of tasks that are intended to replicate on-the-job tasks. A high-fidelity simulation presents job-relevant situations to candidates and requires *behavioral* responses (e.g., work-samples and assessment centers). On the other hand, low-fidelity simulations present a task stimulus using a written or spoken description and elicit only a written or spoken description of the response to the hypothetical scenario (Motowidlo et al., 1990). A casual review of the extant literature suggests that low-fidelity simulations and text/video-based SJTs are synonymous (Weekley, Hawkes, Guenole, & Ployhart, 2015). Compared to high-fidelity simulations (e.g., role-play or work sample), low-fidelity simulations (SJTs) are relatively easy and inexpensive to develop, administer, and score (Clevenger, Pereira, Wiechmann, Schmitt, & Harvey, 2001). However, when additional design features are added to increase their fidelity (e.g., as in video-based SJTs), this results in substantial increases in their cost (Weekley & Jones, 1997).

There are a number of factors and features that come into play in the design and implementation of SJTs. These are interaction modality, fidelity, stem complexity, mode of presentation, response format, scoring key development, scoring strategy, sample-type effects, and response instructions. Generally, much research has been devoted to examining these characteristics because they have been found to influence the validity and other psychometric properties of SJTs.

Interaction modality. *Interaction modality* reflects how interactive (or “branched” or “nested”; Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, & Donovan, 1998) tests are. Thus, SJTs are interactive when the response to one item determines the next item that is presented. Kanning, Grewe, Hollenberg, and Hadouch

(2006) examined how interactivity influences test-takers reactions. Results indicated that participants rated the interactive items as being more “useful” than the non-interactive items.

Fidelity. *Fidelity* is concerned with the degree to which a simulation matches the characteristics of the operational task or environment. While there are multiple dimensions of fidelity, the one that is germane in the context of SJTs is psychological fidelity (Weekley et al., 2015). Psychological fidelity is the extent to which the assessment environment engenders the same psychological processes as the operational environment. Defined this way, one would expect video-based simulations to display higher fidelity than text-based simulations. However, although the video- and text-based distinction can be applied to both the SJT item stem (stem fidelity) and response (response fidelity; Lievens, De Corte, & Westerveld, 2015), it is generally less relevant in terms of the response. This is because in practice, respondents are almost always asked to respond by choosing from a (text) list of pre-specified options. Thus, SJT response fidelity tends to be uniformly low because the task required of the test-taker is to choose a response from a list, as opposed to *performing* a task or watching the performance of a task or behavior. In contrast, and again in practice, there is variability in terms of how the stem is presented—via video or text. Consequently, stem fidelity is typically more variable across studies than response fidelity.

Stem complexity. *Stem complexity* refers to the level of detail used in the item stem (Ployhart & MacKenzie, 2011). For example, simple stems use simple and short language, while more complex stems are longer and usually require more attentional

resources. There is limited research on stem complexity, and it seems that results that have been obtained are not conclusive. For example, McDaniel et al. (2001) found no significant results for their “detail of question” moderator, which speaks to stem complexity.

Mode of presentation. Previous research has examined the effect of *mode of presentation* (i.e., written and video-based) and level of fidelity of SJT on relevant outcomes of interest (e.g., Chan & Schmitt, 1997; Weekley & Jones, 1997). Video-based SJTs typically use video clips to present the SJT stem, after which respondents select a response option. For instance, Chan and Schmitt (1997) examined the effect of mode of presentation (video-based vs. paper-and-pencil) on a variety of outcomes, including subgroup differences, face validity, and SJT test performance. Findings indicated that the African-American/White difference in SJT test performance was substantially greater for paper-and-pencil based SJTs ($d = -0.95$) than video-based SJTs ($d = -0.21$). Furthermore, face validity was higher for video-based SJTs than paper-and-pencil SJTs ($r = -0.24$). Finally, overall, test performance was higher on video-based SJTs than paper-and-pencil SJTs ($d = -0.42$), which is evidence for a mode of presentation effect on SJT test performance.

Richman-Hirsch, Olson-Buchanan, and Drasgow (2000) also examined mode of presentation and found that it can influence test-takers’ reactions. Specifically, there were three versions of the SJT, all with identical linguistic content—a written, paper-and-pencil version, a written form administered by computer, and a multimedia form administered by computer. The multimedia version consisted of a video clip

presentation of situations, while the computerized form was simply the paper-and-pencil form administered via computer. Respondents perceived the multimedia assessment (i.e., with video situations) to be more face valid and reported more positive reactions as well. In summary, the influence of mode of presentation has been examined on a variety of outcomes of interest, and the results indicate that it has an effect on these outcomes.

Response format. SJTs are typically characterized by a multiple-choice item format, a format in which test-takers choose an alternative from a list of response options. *Response format* refers to the way in which respondents are instructed to respond to the SJT items. The three most commonly used response formats in the literature are rate, rank, and most/least. For the rate response format, test-takers rate the effectiveness of each response option as a solution to the scenario. For the rank response format, test-takers rank the response options as a solution to the scenario in order of effectiveness. The most/least response format requires test-takers to only indicate the most effective and least effective response options as solutions to the scenario (it is essentially a truncated version of the rank response format). Test-takers may also be instructed to simply select the best/correct answer. Although all of the preceding response formats are used in the context of a multiple-choice item format, the response format could also entail a constructed-response; that is, an open-ended response in which the participant *produces* an answer (as opposed to being given a list of possible responses).

Previous work has examined the comparative efficacy of different response formats in different contexts (e.g., Arthur, Glaze, Jarrett, White, Schurig, & Taylor,

2014). Specifically, Arthur et al. (2014) found that when designed to measure noncognitive constructs, the rate response format appears to be the superior response format. The use of other response formats (i.e., rank and most/least) may introduce construct-irrelevant variance due to the higher amount of information processing demands engendered by the rank-SJT and most/least-SJT relative to the rate-SJT.

Scoring key development. SJTs are characterized by predetermined scoring keys because the scoring keys are developed a priori. Key development refers to the method for determining the best/correct answer among the response options (Campion, Ployhart, & MacKenzie, 2014). There are two main approaches to key development—the rational-judgmental approach, and the empirical approach. Rational key development (which can be informed by data, such as in the modified Angoff method; Busch & Jaeger, 1990) occurs when subject matter experts (SMEs) make judgments about response options and rate (or rank) the effectiveness of each response as a solution to the scenario. All SME ratings/rankings are aggregated according to a decision rule (e.g., by taking mean, median, or mode; or by reaching a consensus). Empirical key development occurs when each response option is assigned a weight based on its relationship with some criterion of interest (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006) or means/frequencies. Response options with strong correlations are considered correct responses, or at least receive higher weight designations than those with weaker relationships.

Research indicates that there is a wide variability in effectiveness of key development strategies (Bergman et al., 2006). Bergman et al. developed and

demonstrated the use of 11 different scoring approaches based on their review of currently-existing key development and scoring strategies. They found that there was a wide range of validity coefficients depending on the particular combination of scoring strategy and key used (i.e., -.03 to .32). Assessments differed in two main ways, (1) the way in which the correct and incorrect responses were identified (i.e., the key), and (2) the way the key is applied to the responses (i.e., the scoring strategy). Therefore, the combination of the scoring key and scoring method can have a considerable impact on the validity of an SJT.

Scoring strategy. Scoring strategy, that is, how the scoring key is applied to test-takers' responses, has received some research attention. Scoring strategies define the range of scores per SJT item, and can be categorized into two groups. Relative scoring (or partial credit scoring) allows for a range of scores possible for each item within the test, as is the case when the difference (usually absolute) between the key and test-taker's response is calculated for each response option. On the other hand, absolute scoring is dichotomous in nature and allows only for the dual possibility of a correct or incorrect response, as is the case when test-takers' responses (for each response option) are compared to the key and scored in an absolute fashion as being correct or incorrect.

For example, one popular relative scoring strategy for a best/worst response format (i.e., a format in which participants select the best *and* worst response to the scenario) corresponds to the following set of decision rules (Motowidlo et al., 1990). Each of the two alternatives chosen by the participant is given a score of a 0, -1, or 1. Therefore, for each SJT item, participants are given two scores, one for each alternative

they select (best and worst). The alternative chosen as the best response is scored as follows: 1 if it is the keyed best response, -1 if it is the keyed worst response, or 0 if it is neither the best nor worst. The same is done for the worst response selection (i.e., 1, -1, or 0 according to the same rules). Consequently, final item scores can range from -2 to 2. Participants receive a 2 if their best and worst responses match up exactly with the item's best and worst keyed responses. Participants receive a -2 if their best and worst responses match up exactly with the keyed *worst* and *best* responses, respectively.

It is important to acknowledge some recent research regarding scoring methods that has shown evidence for reducing White/African-American mean differences and improving the psychometric properties of the items (McDaniel, Psotka, Legree, Yost, & Weekley, 2011). Specifically, these scoring adjustments are methods that control for elevation (the mean of the items for a given respondent) and scatter (the magnitude of the respondent's score deviations from their own mean). McDaniel et al. suggest that these characteristics primarily reflect response tendencies (e.g., preference for extreme/middle scores) that are a source of systematic error. Therefore, controlling for them reduces error, and this was supported by their findings. Specifically, significant differences were found between criterion-related validities with and without their scoring adjustments. For example, in their Study 1, their raw consensus scoring (i.e., no adjustment) criterion-related validity was .03, while their standardized consensus scoring (i.e., with scoring adjustment) had a statistically different criterion-related validity of .10.

Sample-type effects. The issue of sample-type in the context of SJT research is not a design characteristic or feature per se. Nevertheless, because it has been shown to affect measurement-related outcomes, it is briefly reviewed here. MacKenzie, Ployhart, Weekley, and Ehlers (2009) examined the construct-related validity of SJTs and differences in score means across different samples, specifically applicants and incumbents. They examined construct-related validity and mean differences by comparing correlates of SJTs and various KSAs using applicant and incumbent samples, and their findings indicated that the sample-type is an important moderator of SJT construct-related validity for GMA. Specifically, cognitive ability was more strongly related to incumbent SJT scores than applicant SJT scores. MacKenzie et al.'s (2009) offered explanation was that incumbents with greater cognitive ability can acquire and use firm-specific job knowledge more effectively than applicants. Other results showed that incumbents score higher than applicants on the SJT, with *ds* ranging from .12 to .54. Therefore, there is a sample-type moderator present for SJT mean differences as well.

Response instructions. The current study will focus on one design feature, response instructions. Response instructions can take the form of a would-do instruction set or a should-do instruction set. Would-do instructions ask test-takers to respond in terms of what actually *would* be done in the presented situation, whereas should-do instruction sets ask them to respond in terms of what *should* be done in that situation. For example, for a situation in which the respondent answers within the context of a new hire, a should-do instruction set might read “Rate the effectiveness of each of the

following in terms of what you think a new hire should do.” For the same SJT to have a would-do instruction set, the word “should” is replaced with the word “would.”

Although not explicitly stated in the literature, the present study proposes that the design feature, *response instruction sets*, can be defined in terms of the grammatical stance (not a technical term) of the instruction set (i.e., hypothetical [would-do] or obligatory [should-do]). The next section provides a more detailed review of the literature on this design feature.

Research on Situational Judgment Test Response Instruction Sets

As previously noted, the role and extent of response instructions has been investigated in the literature. Prior work on SJT response instruction sets is reviewed here with particular attention to major findings as well as any relevant limitations. Previous research has found that response instruction sets can affect a variety of outcomes of interest, including the psychometric properties and specifically, construct-related validity of SJTs.

Ployhart and Ehrhart (2003) examined the effect of different types of SJT instructions on the psychometric characteristics of SJTs. Results indicated that the instructions had a large impact on SJT responses, reliability, and validity. Would-do instruction sets were generally associated with more favorable outcomes, including higher criterion-related validities than should-do instructions sets with several criteria such as GPA (*rs* ranging from .30 to .40, depending on the particular version of would-do instructions; corrected *rs* ranging from .41 to .52), peer ratings of academic study skills and behaviors (based on the Learning and Study Strategies Inventory; Weinstein,

1987; r s ranging from .38 to .49; corrected r s ranging from .64 to .68), and self-ratings of academic study skills and behaviors (r s ranging from .47 to .51; corrected r s ranging from .75 to .89). Next, would-do versions of SJTs displayed stronger intercorrelations than should-do versions, which Ployhart and Ehrhart (2003) argue is representative of potential differences in construct-related validity. Average intercorrelations among the would-do versions were high (average uncorrected $r = .76$; average corrected $r = 1.00$), whereas the average intercorrelations among the should-do versions were considerably lower (average uncorrected $r = .36$; average corrected $r = .70$). Ployhart and Ehrhart (2003) also found that would-do versions of SJTs displayed higher test-retest reliabilities than should-do versions. Finally, scale statistics indicated that the should-do versions displayed higher means, smaller standard deviations, and less-normal distributions than their would-do counterparts.

One limitation to the work by Ployhart and Ehrhart (2003) was that changes in response instructions were confounded with changes in response format. For example, should-do instruction sets consisted of “most and least effective,” “rate effectiveness of each response,” and “should do,” whereas would-do instruction sets consisted of “most and least likely do,” “rate how likely would do each response,” and “have done in past”. Figure 1 presents a summary of the variety of response instructions that have been classified under the “would”/“should” nomenclature (McDaniel et al., 2007). Various levels of these response instruction sets are more appropriately considered as types of response format (e.g., most/least, rate) and can be used in either type of response

instruction. Essentially, elements other than word choice of “would” or “should” were also manipulated, which results in lack of isolation of response instruction sets.

	One Scoreable Response	Two Scorable Responses	As Many Scoreable Responses as Response Options
Behavioral Tendency (“would”)	<ul style="list-style-type: none"> ▪ What would you most likely do? 	<ul style="list-style-type: none"> ▪ What would you most likely do? ▪ What would you least likely do? 	<ul style="list-style-type: none"> ▪ Rate each response for the likelihood you would perform the response ▪ Rank the responses from the most likely to the least likely
Knowledge/Ability (“should”)	<ul style="list-style-type: none"> ▪ Pick the best answer ▪ What should you do? 	<ul style="list-style-type: none"> ▪ Pick the best answer and pick the worst answer ▪ Pick the best and second best answer 	<ul style="list-style-type: none"> ▪ Rate each response in terms of its effectiveness ▪ Rank the responses from the best to the worst

Figure 1. Variety of operationalizations of "would-do" and "should-do" instruction sets used in the literature. Summarized from McDaniel et al. (2007).

Faking. Other research indicates that would-do versions are easier to fake than should-do versions. For example, Nguyen, Biderman, and McDaniel (2005) examined faking on SJTs using both a knowledge-based instruction set (corresponding to a should-do instruction set) and a behavioral instruction set (corresponding to a would-do instruction set). Using a within-subjects design, participants were instructed to respond

honestly, and to fake good. Results indicated that the faking effect size for the behavioral tendency response format ranged from .15 to .34 (average $d = 0.25$), depending on order of instructions (i.e., honest first, or fake good first). Therefore, there was a mean difference between honest and faking conditions regardless of order for the behavioral response format. The results of the faking mean difference for the knowledge response format were mixed (average $d = 0.06$). The authors found a positive mean difference for the faking-honest condition, but a negative mean difference for the honest-faking condition. Their rationale for this finding was the following. Participants did their best the first time they took the test. However, the second time, when they were asked to do something different (i.e., to respond honestly if they faked, or to fake if they had responded honestly), participants responded by changing their responses, which means their scores decreased, given they responded to the best of their ability at Time 1. The tentative implication offered by Nguyen et al. (2005) was that SJT scores obtained under the knowledge response format (should-do) could be treated as relatively immune to faking. That is, applicants' scores will be as high as possible regardless of the inclination to fake. This has been considered as support for the proposition by McDaniel et al. (2007) regarding the amount of response distortion in SJTs. Specifically, SJTs with behavioral tendency instructions are considered to be more susceptible to faking (i.e., both unconscious [self-deception] and conscious [impression management]) than SJTs with knowledge instructions.

A related study by Lievens, Sackett, and Buyse (2009) aimed to test the fakeability of SJTs based on response instruction sets in a high-stakes context. Although

most previous work on response instructions was done using incumbents in a low-stakes context, in this experiment, response instructions were manipulated using applicants in a high-stakes context, holding the SJT content constant. Results indicated that there was no difference between responses on SJTs with behavioral tendency instructions and responses on SJTs with knowledge instructions. Therefore, findings regarding mean score differences found in prior research were not replicated in a high-stakes context. Also notable, even in a high-stakes context, the SJTs with knowledge instructions were more correlated with cognitive ability ($r = .19$) than the SJTs with behavioral tendency instructions ($r = .11$). However, this study confounded response instruction sets with response format as well. The two conditions for different response instruction sets were as follows: “pick the best response” (i.e., knowledge or should-do equivalent) and “What would you most likely do?” (i.e., behavioral tendency or would-do equivalent). In addition, although they did hold the content constant, the results cannot speak to differences in construct-related validity for SJTs that are designed to measure different constructs. The present study addresses these concerns.

Construct-related validity. In their meta-analysis, McDaniel et al. (2007) found that response instruction sets influenced the constructs measured by the SJTs. Specifically, tests with should-do instructions had higher correlations with cognitive ability than tests with would-do instructions (i.e., .35 and .19, respectively). The opposite was found regarding relationships with personality traits. Specifically, tests with would-do instructions had higher correlations with personality traits than those with should-do instructions (referred to hereafter as the *instruction set effect*). The higher

correlations between would-do instruction sets than should-do instruction sets and personality traits is clearest for agreeableness (.37 vs. .19; would-do and should-do, respectively), conscientiousness (.34 vs. .24), and emotional stability (.35 vs. .12). Therefore, by examining convergent and discriminant validity, response instruction sets were found to be a moderator of the construct-related validity of SJT scores. One of the key limitations of the McDaniel et al. (2007) meta-analysis is the lack of control for content differences across the SJTs, resulting in a method-construct confound (Arthur & Villado, 2008). The present study addresses this limitation by examining how the effect of response instructions on the construct-related validity of SJTs is influenced by the SJT content.

In addition, there is a conceptual limitation in the McDaniel et al. (2007) meta-analysis as well as most other studies reviewed here. Namely, as briefly mentioned in preceding sections, because the operationalization of response instruction sets has not been standardized, it has typically included changes to other parts of the SJTs. For example, Ployhart and Ehrhart (2003) examined how different response instruction sets influenced the psychometric properties of SJTs; however, their manipulations included changes to *both* response instruction sets and response format. The same can be said about the McDaniel et al. (2007) meta-analysis (see Figure 1). The present study seeks to strictly conceptualize response instruction sets as the grammatical stance of the instructions of the SJT, operationalized as a binary choice between a set of instructions that includes a “would” or “should” (i.e., would-do or should-do instruction sets, respectively).

Operationalizing response instruction sets to be a choice between these two words reflects the dominant thinking about response instruction sets in the literature. Here is one example reflecting how the design feature response instruction sets is typically discussed in the literature, with a subsequent clarification. In McDaniel et al.'s (2007) meta-analysis, the explanation offered for response instruction sets includes a discussion about the distinction between knowledge and behavioral tendency response instruction sets, based on McDaniel and Nguyen (2001). Basically, McDaniel et al. (2007) note that knowledge response instruction sets ask for the correct or best possible response and that behavioral tendency instructions ask for what the respondent would likely do. Although there is no contention here, this kind of language does seem to obfuscate the distinction between knowledge response instruction sets and behavioral tendency instructions. Consequently, it is apparent that the core difference between the two types of response instruction sets is even simpler than the impression the reader might have based on McDaniel et al.'s (2007) discussion.

When respondents are asked for correct or best possible responses (i.e., giving the impression that there is a correct answer), they are essentially being asked what *should* be done in the situation (i.e., implying obligation). In addition, when they are asked for their behavioral tendency/what they would likely do (i.e., giving the impression that there is not necessarily a correct answer), they are being asked what they *would* do in the situation. Therefore, the fundamental distinction between these two types of instruction sets is a question of the grammatical stance of the verb and can adequately be reflected in the word choice between “would” or “should” in the

instructions given. It is also specific enough to allow for a concrete, replicable manipulation of response instruction sets without changing any other part or feature of the SJT (e.g., stem, response options, or response format).

A recent thesis (Kelly, 2013) examined the effect of response instruction sets on SJT test performance in a high-stakes setting. There are numerous notable features of the study. Kelly's manipulation closely approximates the present study's manipulation, specifically the word choice manipulation (i.e., "would" vs. "should"). It is worth noting that this manipulation occurred in two places on the paper-and-pencil SJT—in the overall instructions (at the top of the page) and within the stem of each item. The overall instructions included the focal word in bold. At the end of each stem, respondents were presented with a statement that read "In this situation you would (or should)," while the response options completed the sentence. The content was also held constant (the SJT was designed to measure interpersonal relations). Results indicated there were no significant differences within or between subjects for mean scores by instruction type. In addition, regarding construct-related validity, correlations between SJTs with knowledge instructions and SJTs with behavioral instructions with cognitive ability were .25 and .23, respectively. Based on the information given, the cognitive ability test was a proprietary assessment with coefficient alpha between .88 and .89, depending on the alternate form used. Personality traits were not measured.

Mentioned as a limitation in the thesis, one flaw discussed in the study was that there was no manipulation check. Therefore, it is not clear whether participants processed the differences between the two response instruction sets. In addition,

personality traits were not measured to assess construct-related validity; only cognitive ability was measured. Finally, random assignment was not possible due to practical constraints resulting from the study's field setting. The present study addresses these methodological issues.

Present Study

Despite previous research on the influence of response instructions on the construct-related validity of SJTs, further examination of this issue is warranted. Specifically, as the preceding review suggests, a reading of the literature leaves one to initially conclude that instruction sets influence the construct-related validity of SJTs such that SJTs with a would-do instruction set behave like behavioral measures and SJTs with a should-do instruction set behave like ability measures. However, a more detailed reading and examination of the literature indicates that the instruction set effect is at best only suggestive since it has been based on studies that are methodologically ambiguous.

That is, as indicated in the preceding review of the studies that serve as the basis for this claim, they are characterized by a host of nontrivial methodological concerns. Therefore, the many studies that claim to have examined this issue (e.g., Lievens et al., 2009; Nguyen et al, 2005; Nguyen & McDaniel, 2003; Ployhart & Ehrhart, 2003) do not appropriately do so. Indeed, the one study that seems to have come closest to examining this issue in a methodologically sound manner, Kelly (2013), failed to obtain the instruction set effect. However, because it also failed to include a manipulation check, it remains unclear whether participants processed the particular word (i.e., “would”/“should” manipulation) or not. Also, in other studies (e.g., Lievens et al., 2009;

Nguyen et al., 2005; Nguyen & McDaniel, 2003; Ployhart & Ehrhart, 2003), where response instruction sets have been claimed to be manipulated, they have been confounded because other aspects of the SJT (e.g., other parts of the stem or response format) were concurrently manipulated as well due to the lack of a clear delineation of response instruction sets. Therefore, given the stricter and clearer conceptualization of response instructions offered here, the isolated effect of response instruction sets on the construct-related validity of SJTs has not been examined appropriately thus far. This study seeks to address this fault in the literature.

Because many current claims of the influence of the instruction set on the construct-loading of SJT items speak to a distinction, which, at its core, is a question of this word choice (i.e., “would” vs. “should”), it is important to ground this in empirical findings. In addition, much of the published literature speaks to this issue as a question of word choice. For example, in a recent review, Weekley et al. (2015) discuss that “knowledge instructions (e.g., “What should you do?”) increased racial subgroup differences over behavioral tendency instructions (e.g., “What would you do?”)” (p. 300). Language like this leaves the reader to conclude that manipulations with just the word change are pervasive. Therefore, there seems to be a received doctrine in the literature in the absence of clear, unconfounded research studies that speaks to the effect of the response instruction sets on the construct-related validity of SJTs. To accept this received doctrine, a number of issues need to be addressed and clarified. Specifically, a question remains regarding whether or not the effect of a change in construct-loading of

the SJT (i.e., the instruction set effect) will be present when a single word is changed in the instruction sets (i.e., “would” to “should” and vice-versa).

Consequently, although this claim is widely accepted, there are few empirical studies that have actually examined and/or supported it. In addition, those that have examined it have not been as methodologically sound as they could have been, which calls into question the confidence that can or should be placed on their findings.

As previously discussed, being a method, SJTs have design characteristics that influence their measurement properties, and the design feature of interest to the present study is the response instruction set. Because the instruction set effect (i.e., the finding that would-do SJTs behave like behavioral measures and should-do SJTs behave like ability measures) has not been examined or demonstrated using unambiguous, methodologically sound research designs and protocols, the first objective of this paper is to re-examine this issue in a comprehensive manner. Second, the assertion that instruction sets influence the construct-related validity of SJTs is also examined in the context of the construct that the SJT was designed to measure. The literature on SJT response instruction sets to date has paid very limited attention to the constructs assessed and has instead been primarily method-focused. Because SJTs are methods, an examination of the effects of response instruction sets on the construct-related validity requires considering the construct being measured. The present study addresses this gap in the literature as well.

It is important to note that the present study’s examination has a method-based component (i.e., two types of response instruction sets) *and* a construct-based

component (i.e., two different constructs assessed by the SJT). Therefore, the research design that permits one to effectively speak to the issues at hand consists of the following components: (a) there should be two SJTs, one measuring a cognitive construct and the other measuring a noncognitive construct, and (b) there should be two response instruction sets which are manipulated.

Construct-Related Validity: Would vs. Should

The first objective of the present study is to re-examine whether response instruction sets (i.e., differing only in word choice [“would” vs. “should”]) can change the construct-loading of the exact same SJT. A first step to addressing this research question is to assess whether participants consciously process the word choice at all.

Because no previous studies in the selection literature have included a manipulation check of the single word manipulation in the instruction set of an SJT, it is initially unclear whether or not participants will notice the focal word. However, there are three main frameworks within the cognition literature that can offer some insight, the false memory literature, the anticipation literature, and the eye movements during word processing literature. Relevant work done on false memories is reviewed next.

False memory. The false memory literature has shown numerous times that a subtle word change can have dramatic effects on memory. For example, Loftus and Palmer (1974) had participants watch a video of a car crash and later asked them a number of questions. When asked about whether or not glass was broken in the scene of the car crash, participants responded differently depending on whether the word “smashed” or “bumped” was used. Reading the word “smashed” elicited more false

claims of having seen broken glass in the car accident. This is an example of the misinformation effect, which occurs when exposure to misleading information after an event can lead to false reporting of items and events, and has been shown many times with different manipulations (e.g., Braun & Loftus, 1998; Loftus, 1977). In fact, Loftus and Zanni (1975) showed the misinformation effect occurred with changes as subtle as “a” versus “the.” After participants watched a car crash video, they read and responded to “Did you see *a* broken headlight,” or “Did you see *the* broken headlight?” Findings indicated that presence of the definite article (i.e., “the”) resulted in a greater number of false reports. Therefore, the subtle word change influenced how participants responded to the question. Although the present study does not argue that the instruction set word choice (i.e., “would” vs. “should”) can produce a false memory, the false memory literature is relevant to the present study to the extent that it can provide an explanation for why participants *may* actually perceive the subtle word choice (i.e., “would” versus “should”), which may in turn influence their responses.

Anticipation. The next framework that is relevant to word processing comes from research related to anticipation. It is no surprise that humans are adept at anticipating events in daily life. For example, people may predict that if it is cloudy outside, then it is likely to rain; or that if it is rush hour, then increased traffic is likely. Anticipation also pervades language and language processing. Humans predict upcoming turns in conversations (e.g., Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005) and can anticipate certain information in text while reading (e.g., Wildman & Kling, 1978). Specifically, a number of types of anticipation have been

supported. First, semantic anticipation has been demonstrated in that readers tend to use knowledge of semantic relationships in text to predict semantic characteristics of upcoming words (e.g., Holmes, Arwas, & Garrett, 1977; Olson & McKay, 1974). Second, syntactic anticipation refers to the notion that readers tend to anticipate grammatical structure, and this has been supported as well (e.g., Kollers, 1970; Wanat, 1976). Third, there is also evidence that supports spatial anticipation, which is the reader's anticipation of where the next eye fixation should be in order to maximize information gain (e.g., Shebilske, 1975). It is useful to recognize that expectations or anticipations are a form of top-down processing (as opposed to bottom-up), which happens when readers use prior knowledge and contextual cues to construct meaning (Goodman, 1979).

Given that these types of anticipation have garnered support, it is reasonable to predict that participants will naturally anticipate (whether correctly or incorrectly is a separate, yet empirical question) the focal word (i.e., “would” or “should”) while reading through SJT response instructions, especially given that this processing is primarily top-down. It is easy to speculate that the presence of these forms of anticipation might decrease the level and/or amount of processing of these words, which can in turn limit the amount of meaning derived from these focal words by the reader while taking the SJT. That is, if anticipation influences the depth of word processing at all, then it is likely to influence it in the negative direction, which lowers depth of processing (as opposed to the positive direction, which might increase depth of processing). If true,

then this implies that participants will likely *not* process (i.e., recall) the focal word within SJT instruction sets.

Eye movements during word processing. Finally, another related line of research is the cognition literature on eye movements in word processing. This literature makes a distinction between content and function words (Schmauder, Morris, & Poynor, 2000). Content words are words that add meaning to a sentence (e.g., nouns and verbs), while function words are words that express a grammatical relation or serve a syntactic function (e.g., prepositions). Carpenter and Just (1983) monitored participants' eye movements during reading and found that content words were fixated upon 85% of the time while function words were fixated upon 35% of the time. Furthermore, it is generally held in the cognition literature that eye movements are affected by cognitive processes of readers while reading (e.g., Rayner, 1977; Rayner, 1978). In addition, eye movements from one fixation to the next (i.e., saccades) typically reflect attentional shifts. These saccades serve to bring the next attended feature in the visual array to the central fovea where vision is best (Cunitz & Steinman, 1969).

Given that (1) eye movements reflect attention to words, (2) content words are more likely than function words to be fixated upon, and (3) the present study's focal words are content words (i.e., verbs), this might provide support for a hypothesis stating that participants *will* process the focal words (i.e., operationalized as word recall in the subsequent manipulation check item). However, given that the word manipulation in the present study is "would" versus "should" (i.e., content words), it is entirely possible that these words might be exceptions to the general rule that content words are fixated upon

more than function words. Although findings suggest that content words are fixated upon 85% of the time while function words are fixated upon 35% of the time, this does not mean that every content word is fixated upon. Words like auxiliary verbs (e.g., “should” and “would”) might reasonably contribute to the 15% (i.e., 100% – 85%) of the time that content words are not fixated upon. This seems reasonable when one considers the notion that “would” and “should” are not action verbs like “examine” and “write.” In addition, given the literature on anticipation, it is likely that participants can anticipate the final sentence of SJT response instruction sets being some version of the following idea: “Select a response.”

Therefore, the first hypothesis is that participants will not initially notice the focal word. As shown in Figure 2, which illustrates the conditions used in the present study, responses from all participants in the standard saliency conditions (i.e., cells 1 through 4, or row 1) are relevant here. Furthermore, if this prediction is true and participants do not notice the focal word, then it is conceivable that modifying the focal word (i.e., with **bold/underline**) will increase processing. Regarding Figure 2, this corresponds to the across-row comparison, standard vs. bold/underline (rows 1 and 2, respectively).

It is important to acknowledge the possibility that participants might eventually pay attention to, and process the focal word after a certain number of items for which the instruction is repeated. If this is true, then this would mean that there is an initial number of items for which participants may *not* consciously process the focal word, and

Instruction Set		Would-do			Should-do		
Construct		Cognitive		Agreeableness	Cognitive	Agreeableness	
Saliency	Standard	1A ^a	1B	1C	2	3	4
	Bold/Underline	5			6	7	8

Figure 2. Matrix of 8 study conditions. The numbers in the cells represent the present study's conditions. There are 2 saliency conditions (standard vs. bold/underline), 2 constructs assessed (cognitive vs. agreeableness), and 2 instruction sets (would-do vs. should-do). ^aCell 1 is subdivided into 3 sub-cells and illustrates how each condition contains 3 subgroups that encounter the manipulation check at different checkpoints as follows: 1A (after first and fifth items), 1B (after third and fifth items), and 1C (after fifth item). The same is applied across all cells.

a subsequent number of items for which participants do consciously process the focal word. Because this might be a potential phenomenon during SJT test-taking, a goal of the present study was to also examine *when* participants start to consciously process the response instructions starting at SJT item 1. So, although it is not a major objective, the design of the present study's manipulation check can speak to this. As noted in Figure 2, using a between-subjects design, the manipulation check was administered after different numbers of items have been completed by participants. The details of this approach and the information that was garnered from it are presented in the Method section.

In summary, although the false memory literature provides an explanation for why participants *may* process the focal word, the combination of the anticipation and eye movements literatures provides a more compelling argument for why participants will *not* process the focal word. So, on the basis that (1) participants are likely to anticipate the general content/meaning of the final sentence of the SJT response instructions (i.e., the anticipation literature), and (2) auxiliary verbs might reasonably contribute to the

15% of content words that are not fixated on, (i.e., eye movements during word processing literature) it is posited that:

***Hypothesis 1:** Participants will not notice the focal word (i.e., would or should) in the instruction sets of SJTs such that there will be more participants who do not accurately recall the focal word compared to those who do accurately recall the focal word.*

If participants are not processing the focal word (i.e., “would” or “should”), then it begs the question of whether a stronger saliency (i.e., **bold/underline**) might rectify this. Informal pilot testing was conducted to examine the level of saliency that was likely to increase accurate recall of the focal word. This pilot testing took place in an undergraduate classroom setting and consisted of a slide sequence followed by an informal request for a shows of hands indicating whether students noticed the change. Slides containing one SJT item each were presented to students in sequence. The two SJT items only differed in the focal word of the instruction set. One slide contained a would-SJT item, while the next slide in the sequence contained the should-SJT item (each slide was presented twice in the sequence). When this sequencing was presented without modifications to the saliency of the focal word, there was about a 1% pass rate of students who noticed the change. However, when this was done with either bold or underline, the pass rate increased to about 75% in both cases. Therefore, to examine whether the saliency could impact accurate recall, a stronger manipulation of saliency was deemed more appropriate. Thus, the bold/underline saliency is examined here.

***Hypothesis 2:** A stronger saliency of the focal word (i.e., **bold/underline**) will elicit more accurate recall of the focal word (i.e., would or should) in the SJT instruction sets.*

Due to an absence of manipulating only the instruction word (“would” vs. “should”) in the literature, it remains unclear whether or not the focal word is enough to change the kind of cognitive processing (i.e., knowledge-based vs. behavioral) required by respondents to SJTs to result in the posited construct-change outcomes. However, previous research can offer some guidance. First, would-do instruction sets (i.e., columns 1 and 2 in Figure 2) could be more highly related to personality tests than should-do instruction sets (and vice-versa for should-do instruction sets) for a number of reasons. Namely, would-do SJTs and personality tests call to mind examples of typical behavior; they are both tapping the noncognitive construct domain space, and both can be influenced by social desirability bias (McDaniel et al., 2007). Second, and by the same logic, SJTs with should-do instruction sets (i.e., columns 3 and 4 in Figure 2) could have higher positive correlations with cognitive ability than SJTs with would-do instruction sets because both ask the respondent to judge the maximally correct response. These explanations would support the argument for the widely-purported instruction set effect (i.e., SJTs with would-do instruction sets are more highly correlated with personality traits; SJTs with should-do instruction sets are more highly correlated with GMA). However, in spite of these reasons, if participants are not processing the focal word to begin with (as posited in Hypothesis 1), then it seems unlikely that different focal words will engender different psychological processes during test-taking.

Furthermore, even if participants recall and process the focal word, then it seems unlikely that this subtle word change is substantial enough to, again, fundamentally change the psychological processes engendered by the items. Based on these points, it is posited that:

***Hypothesis 3:** The effect of word choice will not result in the asserted pattern of results that are currently claimed in the literature (i.e., would-do SJTs behave like behavioral measures, and should-do SJTs like ability measures).*

To remain consistent with the extant literature that asserts the instruction set effect, Hypothesis 3 was tested without taking the construct assessed into account.

Construct-Related Validity: Constructs Assessed and Instruction Sets

The second objective of the present study is to examine the extent to which the instruction set effect, if present, is construct-invariant; that is, whether it is influenced by the construct measured by the SJT. Currently, the literature on SJT response instruction sets has paid very limited attention to the constructs that the SJTs in question were designed to measure. Indeed, as reflected in the previously noted limitation of the McDaniel et al. (2007) meta-analysis, a review of the literature indicates that the preponderance of the studies fail to make any mention of, or provide information about the constructs assessed. For example, Chan and Schmitt (2002) mentioned that their SJT was developed based on a job analysis without specifying the construct(s) measured. However, in some later instances, the constructs measured are identified and held constant. For example, Ployhart and Ehrhart (2003), Nguyen et al. (2005), and Lievens et al. (2009) examined the effect of response instruction sets on a number of dependent

variables relevant to SJTs while holding the content constant. Although holding the content constant is a step towards resolving the SJT construct-method confound, it still does not address whether the posited instruction set effect is construct-invariant. To examine this, SJTs measuring two constructs should be included in the study's design (as in the present study).

It is important to examine whether the effect of response instruction sets on the construct-related validity of SJTs is influenced by the constructs assessed because of its implications for SJT design. That is, these findings can speak to the extent to which the content of the SJT further influences the construct-related validity of SJTs (i.e., beyond the effect of just the instruction set). The basis for why the construct assessed is expected to interact with instruction set to influence the construct-related validity of SJTs has to do with construct-irrelevant variance. Ideally, the construct that a test is designed to measure is intended to be the source of the respondents' test scores. However, when factors having nothing to do with the construct assessed systematically influence the respondents' test scores, construct-irrelevant variance is said to be present (Haladyna & Downing, 2004). Examining relationships between test scores and other measures within the construct's nomological network is one primary way to engage in construct validation (Messick, 1995). This includes convergent and discriminant validity (Campbell & Fiske, 1959).

The present study investigates convergent and discriminant validity as part of an SJT's overall construct validation given the construct being assessed. If construct-irrelevant variance can be reduced, then the expected correlation between the SJT scores

and theoretically-relevant measures should be enhanced. Put another way, if more variance in SJT performance is due to the construct being measured (and not other constructs/sources of error), then the correlation between SJT scores and theoretically-relevant variables should be stronger than they would otherwise be. For example, if an SJT is designed to measure a noncognitive construct, then choosing the response instruction that is already responsible for making SJTs more highly correlated with noncognitive constructs than cognitive constructs would seem to enhance the construct-related validity of the SJT. This is because it decreases the noncognitive SJT's construct-irrelevant variance from cognitive constructs.

Given that different response instruction sets are expected to correlate to different extents with different constructs (i.e., should-do instruction sets more strongly with GMA and would-do instruction sets more strongly with personality traits), matching the construct assessed to the instruction set that makes the SJT more highly related to that particular construct will likely decrease construct-irrelevant variance. This is true because what is being measured is already aligned with a construct with which the SJT is more highly correlated, given the instruction set being used. For example, if the variance of a cognitive SJT's test scores that is due to personality traits (which is construct-irrelevant) is higher *because* of the would-do instruction set (as opposed to should-do), then this construct-irrelevant variance could be reduced by changing the would-do instruction set to a should-do instruction set.

Consequently, the second objective of the current study is to examine whether and how the construct assessed interacts with the choice of response instructions to

influence the construct-related validity of SJTs. For example, if there is an instruction set effect on the construct-related validity of SJTs, then this study will examine whether this effect is construct-invariant. Consequently, this involves crossing the content of the SJT with the instruction set (“would” vs. “should”) to examine the extent to which these two characteristics interact to influence the construct-loading of the SJT. As shown in Figure 2, given these four conditions (cognitive SJT/would [column 1], agreeableness SJT/would [column 2], cognitive SJT/should [column 3], and agreeableness SJT/should [column 4]), the relationships between the SJT and the constructs of interest are examined for each condition to address this question.

***Research Question 1:** If participants recall the focal word in the manipulation check item(s), then is the pattern of results found in the present study construct-invariant such that the relationships found between SJT scores (with would-do and should-do instruction sets) and GMA/personality traits (i.e., agreeableness) remain unchanged when different constructs are assessed (cognitive vs. noncognitive constructs)?*

If the widely-purported notion (i.e., the instruction set effect) about the influence of response instruction sets on the construct-related validity of SJTs is found to be true (i.e., participants process the focal word and the instruction set effect is confirmed), then two predictions can be made about its interaction with the content of the SJT. Specifically, if would-do instruction sets are more highly correlated with noncognitive constructs, then this relationship should be strengthened in cases where the SJT is designed to measure personality traits (i.e., column 2 in Figure 2). That is, if both

personality-based measures and SJTs with would-do instruction sets prime behavioral tendency responses, then the combination of these two elements should more strongly prime the same behavioral tendency responses. A stronger relationship is expected here because the use of the instruction set more highly correlated with the construct of interest is likely to reduce construct-irrelevant variance in the responses compared to an instruction set less correlated with the construct of interest. If instruction sets are correlated with constructs other than those that are intended to be measured, then this will likely introduce construct-irrelevant variance. By using instruction sets that already correlate with the construct of interest, the construct-irrelevant variance, and consequently contamination, is being minimized (i.e., at least compared to using the other instruction set). Furthermore, and for the same reason, if both ability-oriented SJTs and should-do instruction sets prime knowledge-based responses, then the combination of these two elements (i.e., column 3 in Figure 2) should more strongly prime the same knowledge-based responses. Therefore, if the alleged instruction set effect is true, then a number of predictions can be made.

***Hypothesis 4:** Cognitive SJTs with should-do instruction sets will be more highly correlated with GMA than cognitive SJTs with would-do instruction sets.*

***Hypothesis 5:** Noncognitive SJTs with would-do instruction sets will be more highly correlated with personality traits (i.e., agreeableness) than noncognitive SJTs with should-do instruction sets.*

METHOD

Participants

To estimate the sample size required to detect the hypothesized effects, a power analysis was conducted using G*Power 3.1 (Erdfeider, Faul, & Buchner, 1996; Faul, Erdfeider, Buchnar, & Lang, 2009). For the present study, the effect size of interest (Cohen's q) represents the difference in standardized correlations between an SJT with a would-do instruction set and an SJT with a should-do instruction set with other scores of interest (GMA/personality measure). That is, Cohen's q represents the difference between two Fisher-z-transformed correlations (Faul et al., 2009). To detect a small-to-medium effect ($q = 0.17$) with a power of .80, a sample size of 834 was required. Given this effect size, the present study consists of a sample of 264 participants, which translates into a power of .40 to detect a small-to-medium effect ($q = 0.17$). The implications of this particular sample size and associated levels of power for the study's results and conclusions are addressed in the Discussion. Recruited from the Texas A&M Department of Psychological and Brain Science's subject pool, participants' mean age was 19.99 years ($SD = 1.18$), and 59.80% of the sample was female.

Measures

General mental ability (GMA). The GMA test used in the study is an online speeded, 4-alternative multiple-choice test with 60 items (36 verbal and 24 quantitative; Arthur, 2017b). Participants were given 10 minutes to complete the test, and scores were calculated as the total number of items completed correctly. Naber, Arthur,

Edwards, and Franco-Watkins (2017) reported a 7-to-10 day retest reliability of .76 for the scores obtained from this measure.

Agreeableness. Agreeableness was the noncognitive construct and was measured using the 10 items from the International Personality Item Pool (IPIP) measure of the five-factor model of personality (Goldberg, 1999). Participants responded to items using a 5-point Likert scale (1 = very inaccurate; 5 = very accurate) and were asked to indicate the extent to which the given item statements are descriptive of themselves. An internal consistency reliability estimate of .76 was obtained for the present study.

Cognitive SJT. Two SJTs were administered in the present study, a cognitive SJT and a noncognitive SJT. In the *cognitive* SJT, participants were presented with five scenarios in which they are stranded in harsh conditions (e.g., the jungle) with four objects (e.g., flares, water canteen, food, parachute). Each object corresponds to a response option (i.e., there are four response options). Participants are instructed to think about the salvaged objects in terms of which one they would need most to survive and rank-order the items (without ties) in terms of how important they are for survival. The response scale for the rankings ranged from 1 (most important for survival) to 4 (least important for survival). Responses were scored using an absolute scoring approach. Each SJT item has 4 sub-items, each corresponding to a response option. If a correct response (i.e., rank-order) was given for a particular sub-item, then that sub-item was counted as correct. Therefore, there was a total of 20 possible points for the SJT (i.e., 5 items, 4 response options [sub-items] each). Please see Appendix A for a sample

cognitive SJT item. An internal consistency reliability estimate of .86 was obtained for the cognitive SJT scores.

In developing the cognitive SJT, the initial version consisted of five items adapted from the following survival exercises, *Space Survival Exercise*, *Lost at Sea* (Knox, 2009), *Jungle Survival Situation: Leader's Guide* (Lafferty, 1987), *Stranded in the Desert* (Johnson & Johnson, 1991) and *Winter Survival Exercise*. These survival exercises are decision-making tasks that have been used as team-building and other group activities, amongst others. However, they were modified for the purposes of this study to serve as the basis for the cognitive SJT. The SJT development and review process is described next.

First, a search for publicly available, standardized survival exercises was undertaken and resulted in the location of eight exercises. From this initial list, five were retained on the basis that they were reasonably unique or different from each other, based on SME judgment. SMEs consisted of a team of three upper-level students in an I/O psychology Ph.D. program led by a professor of I/O psychology. These SMEs examined each survival exercise for critical elements and created a self-contained, condensed version of its scenario. This was done using an iterative, conjointed review process that involved all team members. The end product of this stage was a set of five condensed scenarios that served as the SJT stems.

The next step was to develop the SJT responses which were based on the keys for the original survival exercises. The keys for the survival exercises that were identified and retained all consisted of a list of responses, ranging from 12 to 15, which were

ranked in importance to survival. (These responses comprised items or objects [e.g., flares, water canteen, food, parachute].) Thus, each exercise that was retained had a predetermined list of rank-ordered responses in terms of their importance to survival in the specified scenario. Because the intent was to design 4-alternative SJT items, the next step was to reduce the 12-15 ranked responses from the original response lists to 5 (which was later reduced to 4). Using the 15-response list as an example, the initial step was to create lists of 5 responses each by selecting every other third response. Thus, the first list consisted of ranked responses 1, 4, 7, 10, and 13; and the second list 2, 5, 8, 11, and 14, etc. The development team then met to review these subsets of responses and through extensive discussion and an iterative review, settled on a final draft of five responses (which was later slightly modified to contain 4 responses) that retained the initial relative rank order but also covered the range of initial responses. So, for instance, for the space survival exercise item, which had an original list of 15 responses, the final list of 5 responses had rank-ordered positions of 1, 4, 8, 10, and 15 on the original list. As mentioned previously, this final draft was later modified to contain only four response options, in order to contain the same number of response options as the noncognitive SJT used in the present study. To accomplish this, the lowest-ranked item was dropped as a response option. For example, within the same space survival item containing originally-ranked positions of 1, 4, 8, 10, and 15, the item located at position 15 was dropped. The same approach was used to create the 4-alternative SJT responses for all five items. Finally, each response also included a behavioral description about

how the item was to be used for survival, adapted from the original scenario key (see Appendix A for an example).

The cognitive SJT is deemed as such because it is a problem solving and decision-making task. Essentially, participants read scenarios that contain a problem or dilemma and decide about the effectiveness of each response as a solution to the problem. First, because it can be argued that these adapted SJTs have a problem-solving component, it is useful to note that the meta-analyzed relationship between complex problem solving and intelligence is a mean Hedge's g of .43, with r 's ranging from -.30 to .86 (Stadler, Becker, Gödker, Leutner, & Grieff, 2015). Second, research has consistently demonstrated that decision-making tasks, such as those used here, are correlated with GMA. LePine, Colquitt, and Erez (2000) report that the relationship between decision-making performance and GMA is .23. They presented participants with decision-making tasks whose rules (i.e., to solve a given scenario) would change periodically, giving the participants the opportunity to adapt. LePine et al. found that the relationship between GMA and performance increased from .23 to .43 after the change. While this increase in GMA loading speaks to adaptability, the initial relationship and overall trend is that GMA is used in decision-making tasks.

Day, Arthur, Miyashiro, Edwards, Tubré, and Tubré (2004) used the desert survival task—one of the tasks used here—and at the team level, found it to have a small but significant correlation with GMA (.18). Finally, Ang et al. (2007) also obtained a correlation of .17 between their measure of cultural judgment and decision making and GMA. Their measure used five cross-cultural decision making scenarios in which

participants were asked to decide which response option best explained the cultural interaction presented in the stem. In summary, illustrative of, and consonant with their cognitive nature, decision-making tasks are reported in the literature to display small to moderate relationships with cognitive ability.

Agreeableness (noncognitive) SJT. The noncognitive SJT is a 5-item measure of agreeableness (Arthur, 2017a). For each item, participants were presented with a scenario and four possible responses to the scenario. Participants ranked the effectiveness (1 = most effective; 4 = least effective) of each alternative (without ties) as a response to the scenario. Responses were scored using an absolute scoring approach. The present study obtained an internal consistency reliability estimate of .91 for the agreeableness SJT scores. See Appendix A for a sample agreeableness SJT item¹.

Manipulation check item. Finally, to determine whether participants processed the focal word (i.e., operationalized as word recall), a manipulation check (MC) item was included as follows: “Below is the instruction prompt for the item(s) to which you just responded. Please fill in the blank with the word that was used.” The focal word in the SJT response instructions was replaced with a blank. Thus, the MC item required a constructed response. A correct response was a match between the focal word given in the instruction set and the participant’s constructed response. There were no spelling errors in any of the constructed responses collected (e.g., “shoud” or “wiuld”).

¹ Due to the proprietary nature of the present study’s items, the sample agreeableness SJT item provided in Appendix A is not one of the five items administered in the present study.

As noted in Figure 2, the MC item had one of three placements (to which participants were randomly assigned) between subjects within the SJT, specifically (1) after the first and fifth items, (2) after the third and fifth items, and (3) after the fifth item. Figure 3 presents an illustration of the sequencing of the various MCs with SJT items based on their location. This MC strategy allowed for a more comprehensive examination of the processing of the focal word and accomplished a number of goals. First, this MC approach allowed for an examination of the repetition effect (if any) on the processing of the focal word. The repetition effect refers to the notion that participants may process the focal word after having been exposed to it (i.e., through reading the instruction set) a certain number of times. Placing the MC at various points in the SJT (i.e., after the first, third, and fifth items) between subjects allows for an examination of whether the amount of exposure, as reflected in the number of items completed prior to the first MC encountered, is related to the processing of focal word. Thus, comparisons were made between participants who responded to the MCs at each checkpoint (i.e., after the first, third, and fifth items) to address this.

Second, the MC strategy allowed for an examination of base rates of focal word recall after the first item, which pertains to Hypothesis 1. One goal of the present study was to ultimately examine how to elicit processing of the focal word at the SJT's first item. Comparisons could be made between the post-first item MC across conditions. Third, the MC approach also allowed for an examination of the priming effects of the MC item. Comparisons could be made between participants who respond post-fifth item as their first MC and participants who respond post-fifth item as their second MC.

MC Location 1	MC Location 3	MC Location 5
<ul style="list-style-type: none"> ▪ SJT item 1 <ul style="list-style-type: none"> ○ Post-1st MC^a ▪ SJT item 2 ▪ SJT item 3 ▪ SJT item 4 ▪ SJT item 5 <ul style="list-style-type: none"> ○ Post-5th MC^b 	<ul style="list-style-type: none"> ▪ SJT item 1 ▪ SJT item 2 ▪ SJT item 3 <ul style="list-style-type: none"> ○ Post-3rd MC^a ▪ SJT item 4 ▪ SJT item 5 <ul style="list-style-type: none"> ○ Post-5th MC^b 	<ul style="list-style-type: none"> ▪ SJT item 1 ▪ SJT item 2 ▪ SJT item 3 ▪ SJT item 4 ▪ SJT item 5 <ul style="list-style-type: none"> ○ Post-5th MC^a

Figure 3. Sequence of SJT items and manipulation checks presented to participants. Depending on the condition to which they were randomly assigned, participants' completion of the MC item corresponded to one of three sequences. It should be noted that all participants received the post-5th MC. ^aDenotes a first-encounter MC. ^bDenotes a second-encounter MC. First-encounter MC items refer to MC items that are administered first in the SJT, regardless of location. Second-encounter MC items, which are always post-5th item MCs, refer to MC items that are administered second in the SJT, and are thus present in only SJTs with MC locations 1 and 3.

Finally, placing the MC at the end of the SJT (i.e., after the fifth item) permitted the examination of whether participants processed the focal word at all by the end of the SJT. Another goal of the present study was to examine the relationships between SJT scores and the constructs of interest when participants do *not* process the focal word. The present MC strategy provided information germane to this goal as well.

Demographics. Participants provided basic information such as age, sex, and race/ethnicity.

Design and Procedure

The study was a 2 (response instructions: would vs. should) × 2 (focal word saliency: standard vs. **bold/underline**) × 2 (construct assessed: cognitive vs.

agreeableness) between-subjects factorial design where the SJT served as the primary independent variable, with 8 levels. Dependent variables consisted of the manipulation pass rates and the specified intercorrelations of interest (e.g., the correlation between the would-do SJT and agreeableness scores, and the correlation between the should-do SJT and GMA scores).

There were eight between-subjects conditions, one for each variant of the SJT. To create each condition, SJTs were manipulated along three factors, construct assessed, the instruction set, and the saliency of the instruction set. The *construct assessed* by the SJT was either GMA or agreeableness. The *instruction set* of the SJT was either a would-do or should-do instruction set. Finally, the *saliency of the instruction set* refers to modifications (or lack thereof) to the word “would” or “should” within the stem of each item. There were 2 levels of saliency, standard and **bold/underline**. In the standard version, there were no modifications to the text of the focal instruction set word (i.e., “would” or “should”). This was the baseline condition. In the **bold/underline** condition, the focal word was both bolded and underlined. Regarding MC location, there were three locations, after the first and fifth items, after the third and fifth items, and only after the fifth item.

The experimental protocol consisted of large-scale (i.e., in a large classroom), proctored Internet testing sessions. Participants first chose to participate in the study by signing up for it through the subject pool recruitment system, at which time they were instructed to bring a laptop to the experimental session. After signing up, participants were given a date to come to a large classroom on campus. Before they arrived for the

experiment, they were also sent a uniform resource locator (URL; also termed “web address”) to the assessment that corresponded to their randomly-assigned condition. This link was activated once the experimental session began.

Once they arrived for the experiment, they began the study by clicking the URL they were sent prior to arrival. This link was *only* active during the timeframe of the experiment. The first page they read was the online consent form, which explained the purpose and instructions of the study. After they read and affirmed the consent form, participants began their individual assessment, which included the following measures in the following order, the online GMA measure, the IPIP agreeableness measure, an SJT (one of 8 versions) with the MC placed in one of three locations (to which participants were randomly assigned), and a demographics measure. That is, they completed one of 8 different SJTs (each corresponding to a unique URL), along with other measures. Therefore, depending on their particular condition, participants completed an SJT that was either cognitive or noncognitive, used either a would-do or should-do instruction set, with one of two different types of saliencies (i.e., types of presentation of the word “would” or “should”), and with a MC placed in one of three predetermined locations. The GMA test was timed such that once the time limit was reached, the test ended and answers up to that point were recorded and saved. Participants completed the other measures at their own pace. Excluding four participants whose time to completion exceeded a day (likely due to lack of officially submitting the survey during the session), the average time to completion was about 36 minutes ($M = 36.06$, $SD = 9.47$). Therefore, no participant exceeded the allotted two hour time stop limit.

RESULTS

Descriptive statistics and correlations of all study variables are provided in Table 1. There were 264 participants total. Of those, 135 took the agreeableness SJT, and 129 the cognitive SJT. Overall, the intercorrelations between major study variables were all in the anticipated direction. This is important because it confirms that the cognitive SJT that was developed for the purposes of this study displays relationships in alignment with its construct assessed. Namely, there was a significant relationship between the cognitive SJT and cognitive ability ($r = .20, p < .05$). It also does not correlate with constructs that are not expected to correlate with it (e.g., agreeableness). Furthermore, there was a significant relationship between the agreeableness SJT and IPIP agreeableness ($r = .26, p < .01$), which provides support for the agreeableness SJT's construct-related validity as well. All analyses were ultimately based on raw scores. However, when appropriate for the purposes of interpretation, these were converted into percentages or z-scores.

Hypothesis Testing

Hypothesis 1 stated that “*Participants will not notice the focal word (i.e., would or should) in the instruction sets of SJTs such that there will be more participants who do not accurately recall the focal word compared to those who do accurately recall the focal word.*” To test this hypothesis, a chi-square test was conducted to examine the pass rate of the MC item for correct identification of the response instruction set. The pass rate was operationalized as the number of participants who could correctly identify

Table 1. Descriptive Statistics and Intercorrelations for All Study Variables

	<i>M</i>	<i>SD</i>	<i>n</i>	1	2	3	4	5	6	7	8
1. Cognitive Ability	32.80	7.12	264	-							
2. Agreeableness IPIP	3.99	0.52	264	.08	-						
3. Agreeableness SJT overall	51.63	18.97	135	.04	.26*	-					
4. Cognitive SJT overall	32.44	12.06	129	.20*	.07	-	-				
5. Agreeableness SJT (would)	50.92	18.22	65	.02	.20	-	-	-			
6. Agreeableness SJT (should)	52.29	19.26	70	.12	.28	-	-	-	-		
7. Cognitive SJT (would)	32.06	12.75	63	.04	-.16	-	-	-	-	-	
8. Cognitive SJT (should)	32.80	11.44	66	.24	.24	-	-	-	-	-	-

Note. * $p < .05$, one-tailed. There is no reported correlation between the cognitive SJT and the agreeableness SJT because these variables are between-subjects. There is also no reported correlation between SJTs of different instruction sets with each other because these variables are between-subjects as well.

the response instruction set in the MC. Thus, the pass rate of a particular MC refers to the percent of participants in a given condition who passed that MC item. The frequencies of pass rates by saliency condition as well as the results of the chi-squared tests are provided in Table 2. Pass rates ranged from as low as 27.3% to as high as 90.7%, depending on various condition characteristics (e.g., MC location, construct assessed, and saliency). Rates of interest in examining Hypothesis 1 correspond to rates in the standard saliency condition, as this is how participants generally encounter the focal words of interest in practice (i.e., without bold or underline modifications).

Correct recall of a particular MC is operationally defined as having a pass rate statistically higher than the comparison rate of 50%, which is considered here to be a conservative comparison rate. For the purposes of the present study, a conservative test

is one that errs on the side of caution and likely reduces the Type 1 error rate as a byproduct of this caution. Fifty percent is the comparison rate of choice in the present

Table 2. Manipulation Check Pass Rates and Associated Chi Square Tests

Manipulation Check Location	Pass rate (percent passing)	<i>n</i>	χ^2	Φ
<i>Standard Conditions</i>				
1	36.4	44	3.27	0.27
3	45.5	44	0.36	0.09
5A	73.3	131	28.40*	0.47
5B	51.1	43	0.02	0.02
5C	84.1	88	40.91*	0.68
<i>Bold/Underline Conditions</i>				
1	53.5	43	0.21	0.07
3	63.0	46	3.13	0.26
5A	77.4	133	40.07*	0.55
5B	52.3	44	0.09	0.05
5C	89.9	88	58.91*	0.82

Note. 5A includes all participants who took this manipulation check item, regardless of condition. 5B includes *only* participants who saw the post-5th MC item. 5C includes *only* participants who saw the 5th MC item for a second time.

study for two reasons. First, proponents of the instruction set effect would likely propose that almost 100%, if not 100% of test-takers process the focal word. However, a pass rate that is considerably lower than 100% can still be statistically higher than 50%; therefore, this is considered to be a conservative test of Hypothesis 1. Second, this comparison rate is also appropriate when one considers the possibility of guessing correctly. In the most conservative case, 50% represents the probability of guessing correctly when there are only two response instruction words in the universe (i.e.,

“would” and “should”). However, there are conceivably more than two words that participants might guess in addition to these two words (i.e., “could” and “might”). Therefore, this relatively high comparison rate, which could arguably be lowered due to the existence of more than two appropriate instruction set words, also makes this test a conservative test of Hypothesis 1.

To address Hypothesis 1, the standard saliency condition MC post-1st item, which is the condition that most closely answers the question regarding whether test-takers notice the focal word from the beginning (i.e., item 1) of an SJT, was examined. These frequencies best represent the participants’ “initial” noticing of the word because it does not include the effects of priming (i.e., that is expected to occur during the time of a *second-encounter* of the MC, if there was one). For the sake of completeness, the pass rates at all MC locations are also reported in Table 2.

The pattern of results indicates that participants generally do not recall the focal word, with the exception of a few of the post-fifth item MC conditions. Figure 4 illustrates pass rates across MC locations. For example, for the standard MC item 1, only 36.4% of participants passed, and this is statistically not different from the 50% comparison rate, $\chi^2_{(1)} = 3.27, p > .05$ ($\phi = .27$). For the standard condition, first-encounter MC item 5 (column 5B in Figure 4), the pass rate was 51.1%, and this is statistically not different from 50%, $\chi^2_{(1)} = .02, p > .05$ ($\phi = .02$). First-encounter MC items refer to MC items that participants encounter first in their assessment, regardless of location (post-1st item MC, post-3rd item MC, or post-5th item MC; see Figure 3 for an illustration of the particular sequencing of items and MCs). Second-encounter MC items

refer to MC items that participants encountered second in their assessment, if there was one.

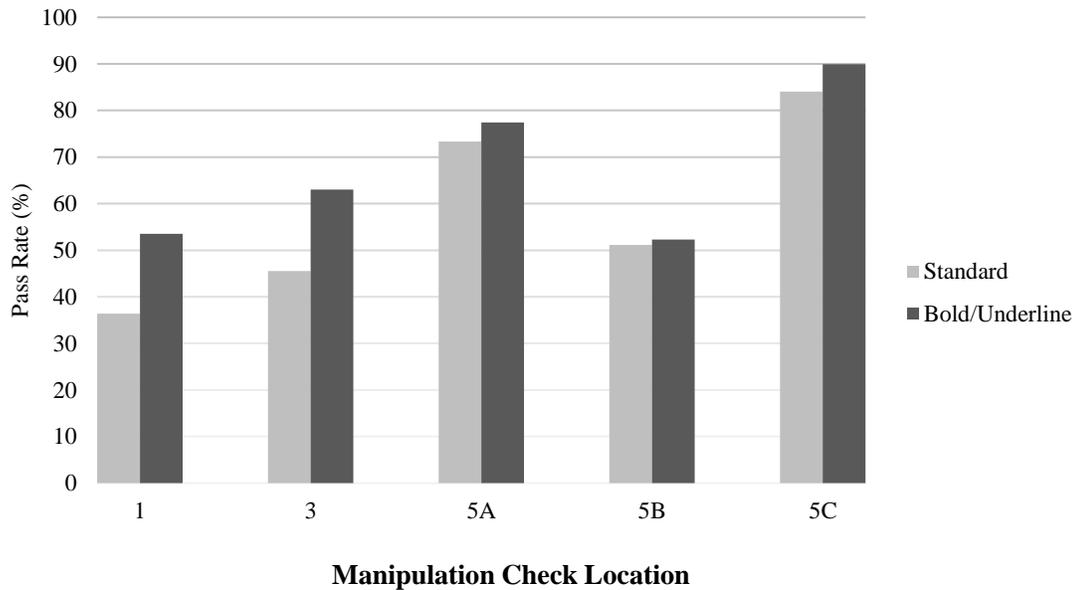


Figure 4. Manipulation check pass rates by condition. 5A represents all participants who took post-5th MC item, regardless of condition. 5B represents the pass rate of the 5th MC item for participants who were administered *only* the post-5th MC item. 5C represents *only* participants who were administered the post-5th MC item a second time (i.e., *only* those participants who were administered two MCs).

Both of these pass rates are no different than the comparison base rate, which suggests that participants are not processing the word in these instances. The same trend is typically found in cases for which the MC pass rate was for a *first-encounter* MC. However, in instances when participants saw the MC item as their second-encounter MC, they tended to do better. For example, for the standard MC 5 (second-encounter; column 5C in Figure 4), 84.1% of participants passed, and this is statistically different

from 50%, $\chi^2_{(1)} = 40.91, p < .05$ ($\phi = .68$). For the sake of completeness, the results illustrated in Figure 4 are presented in Figure 5 broken down by the construct assessed. In general, a similar pattern and magnitude of results were obtained.

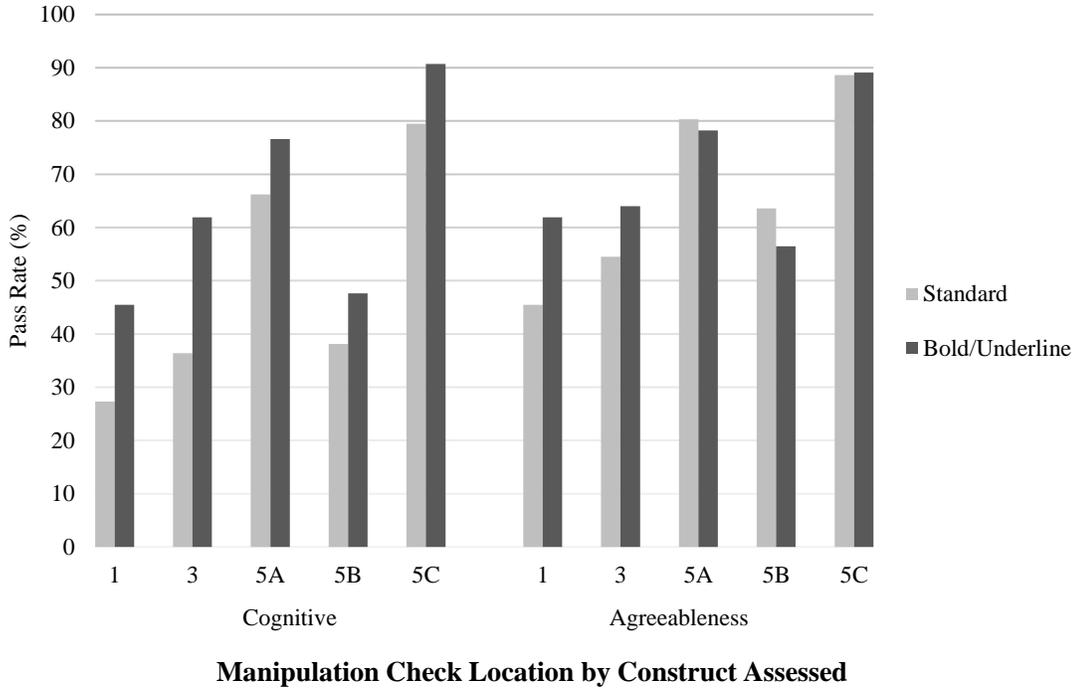


Figure 5. Manipulation check pass rates by location and construct assessed. 5A represents all participants who took post-5th MC item, regardless of condition. 5B represents the pass rate of the 5th MC item for participants who were administered *only* the post-5th MC item. 5C represents *only* participants who were administered the post-5th MC item a second time (i.e., *only* those participants who were administered two MCs).

Overall, except for some instances where the initial MC may have primed participants to pay attention to the focal word, these results indicate that participants do not recall the focal word any more than the default comparison rate (50%), and in some

instances recall the focal word considerably less than the default “non-pass” rate (i.e., MC item 1, in Figure 4). Furthermore, less than half of participants recalled the focal word at item 1. Taken together, these findings support Hypothesis 1; participants do *not* tend to accurately recall the focal word under normal testing circumstances (standard saliency).

As can be seen from a glance at the pass rates in Table 2, the pattern of results suggests that the pass rate is lower during earlier MCs than later ones. Therefore, it can be reasonably concluded that a repetition effect is present in the recall of the focal word. The pass rate for the MC in the standard saliency condition after the 1st item was 36.4%, while the post-5th item MC was 51.1% (first-encounter). Therefore, all other things equal, the rate increased as a function of the number of items that were presented before the first-encounter MC. Again, for the sake of completeness, Table 3 displays the same results broken down by the construct assessed.

Hypothesis 2 stated that “A *stronger saliency of the focal word (i.e., **bold/underline**) will elicit more accurate recall of the focal word (i.e., would or should) in the SJT instruction sets.*” To test this question, a chi-square test was conducted to compare pass rates for correct identification of the response instruction set (i.e., pass rates) between the standard condition and the non-standard saliency condition (i.e., **bold/underline**). See Table 4 (and illustrated in Figure 4) for results including pass rates, chi-squared values, and phi coefficients. As these results indicate, although the pattern of results was generally in the expected direction with the pass rates for the **bold/underline** condition being consistently higher than the standard conditions, none of

Table 3. Manipulation Check Pass Rates and Associated Chi Square Tests

Manipulation Check Location	SJT	Pass rate (percent passing)	<i>n</i>	χ^2	Φ
<i>Standard Conditions</i>					
1	Cognitive	27.3	22	4.55*	.45
3	Cognitive	36.4	22	1.64	.27
5A	Cognitive	66.2	65	6.79*	.32
5B	Cognitive	38.1	21	1.19	.24
5C	Cognitive	79.5	44	15.36*	.59
1	Agreeableness	45.5	22	0.18	.09
3	Agreeableness	54.5	22	0.18	.09
5A	Agreeableness	80.3	66	24.24*	.61
5B	Agreeableness	63.6	22	1.64	.27
5C	Agreeableness	88.6	44	26.27*	.77
<i>Bold/Underline Conditions</i>					
1	Cognitive	45.5	22	0.18	.09
3	Cognitive	61.9	21	1.19	.24
5A	Cognitive	76.6	64	18.06*	.53
5B	Cognitive	47.6	21	.05	.05
5C	Cognitive	90.7	42	30.86*	.86
1	Agreeableness	61.9	21	1.19	.24
3	Agreeableness	64.0	25	1.96	.28
5A	Agreeableness	78.2	69	22.04*	.56
5B	Agreeableness	56.5	23	.39	.13
5C	Agreeableness	89.1	46	28.17*	.78

Note. 5A includes all participants who took this manipulation check item, regardless of condition. 5B includes *only* participants who saw the post-5th MC item. 5C includes *only* participants who saw the 5th MC item for a second time.

the differences were statistically significant. This is not inconsistent with the relatively small magnitude of the differences ($\Phi = .01-.26$). It had been expected that the **bold/underline** manipulation would result in higher recall rates. However, because

there were no significant differences between the two saliency conditions, Hypothesis 2 was not supported. (For the sake of completeness, results illustrated in Table 4 are presented in Table 5 broken down by the construct assessed. In general, a similar pattern and magnitude of results were obtained.)

Table 4. Pass Rate Comparisons between Standard Saliency and Bold/Underline Conditions

Manipulation Check Location	Percent passing (Standard)	<i>n</i> Standard	Percent Passing (B/U ^a)	<i>n</i> B/U	χ^2	ϕ
1	36.4	44	53.5	43	2.58	0.17
3	45.5	44	63.0	46	2.81	0.18
5A	73.3	131	77.4	133	0.62	0.05
5B	51.1	43	52.3	44	0.01	0.01
5C	84.1	88	89.9	88	1.87	0.10

Note. ^aB/U signifies Bold/Underline.

Hypothesis 3 stated that “*The effect of word choice will not result in the asserted pattern of results that are currently claimed in the literature (i.e., would-do SJTs behave like behavioral measures, and should-do SJTs like ability measures).*” Consequently, to test this hypothesis, correlations were computed between the would-do or should-do instruction set SJT scores (collapsing across construct assessed) and the constructs of interest (i.e., GMA and IPIP agreeableness scores). The goal of this analytic strategy was to examine whether the instruction set effect currently purported in the literature could be replicated under the present study’s focal word manipulation. To remain consistent with the extant literature examining this issue, the construct assessed was not

Table 5. Pass Rate Comparisons between Standard Saliency and Bold/Underline Conditions by Construct

Manipulation Check Location	SJT	Percent passing (Standard)	<i>n</i> Standard	Percent Passing (B/U^a)	<i>n</i> B/U	χ^2	ϕ
1	Cognitive	27.3	22	45.5	22	1.57	.19
3	Cognitive	36.4	22	61.9	21	2.81	.26
5A	Cognitive	66.2	65	76.6	64	1.71	.12
5B	Cognitive	38.1	21	47.6	21	0.39	.10
5C	Cognitive	79.5	44	90.7	43	3.17	.19
1	Agreeableness	45.5	22	61.9	21	1.17	.16
3	Agreeableness	54.5	22	64.0	25	0.43	.10
5A	Agreeableness	80.3	66	78.3	69	0.09	.03
5B	Agreeableness	63.6	22	56.5	23	0.28	.08
5C	Agreeableness	88.6	44	89.1	46	0.01	.01

Note. 5A includes all participants who took this manipulation check item, regardless of condition. 5B includes *only* participants who saw the post-5th MC item. 5C includes *only* participants who saw the 5th MC item for a second time. ^aB/U signifies Bold/Underline.

taken into account. Namely, this analytic strategy tested whether would-do SJT scores correlate more strongly with personality traits (i.e., agreeableness) than should-do SJT scores, and likewise whether should-do SJT scores correlate more strongly with the GMA measure than would-do SJT scores. A z-test employing Fisher's r-to-z transformation was used to examine whether correlations were statistically different. In addition, to complement this analysis, a moderated regression was computed as well, with the instruction set as the moderator. The predictor was the construct of interest (i.e., GMA or agreeableness) and the outcome was SJT scores. If a moderator effect is present, then the interaction term would have a significant effect on the outcome (Aguinis, Beaty, Boik, & Pierce, 2005).

The pattern of results did not replicate the effect currently purported in the literature regarding the effect of instruction sets on the construct-related validity of SJTs. First, collapsing across constructs, the relationship between GMA and would-SJT scores was negligible ($r = -.01, p > .05$). Second, the relationship between agreeableness and would-SJT scores was also negligible ($r = .05, p > .05$). Third, the relationship between GMA and should-SJT scores was moderate and significant ($r = .22, p < .05$). And finally, the relationship between agreeableness and should-SJT scores was also moderate and significant ($r = .22, p < .05$). A summary of these relationships can be found in Table 6. Furthermore, a z-test employing Fisher's r-to-z transformation indicates that the GMA/would-SJT correlation and the GMA/should-SJT correlation were statistically different ($p < .05$), while the agreeableness/would-SJT correlation and the agreeableness/should-SJT correlation were not different ($p > .05$).

Table 6. Hypothesis 3: Correlations between Would-do and Should-do SJT scores and GMA and Agreeableness

	Would SJT scores	Should SJT scores
GMA	↓ -.01 ^a ($n = 128$)	↑ .22* ^a ($n = 136$)
Agreeableness	↑ .05 ^b ($n = 128$)	↓ .22* ^b ($n = 136$)

Note. ^aCorrelations are significantly different ($p < .05$) ^bCorrelations are not significantly different ($p > .05$). * $p < .05$, one-tailed. The vertical arrows represent the posited effects such that ↑ represents an expected stronger relationship and ↓ a weaker one.

It is conceivable that the instruction set effect could be masked by the inclusion of participants who did and did not recall the focal word of interest. Therefore, to further examine whether an instruction set effect could be replicated, the same analyses

described above were conducted excluding participants who did not pass the post-5th item MC. It is important to acknowledge that the ideal sample of “passing” participants would be *only* those participants who passed MC item 1 (i.e., because these participants are the only ones who surely processed the focal word at the beginning of the SJT, and subsequently thereafter). However, given sample size concerns, a more relaxed threshold was used; thus, any participant who passed the post-5th item MC was included. While this operationalization does not include *only* people who recalled the word, it does exclude people who did *not* recall the word. Therefore, at least, it was a step in the right direction.

A pattern of results similar to those of the prior set of analyses was obtained. A summary of these relationships can be found in Table 7. The relationships between both GMA ($r = -.03, p > .05$) and agreeableness ($r = .04, p > .05$) and would-SJT scores were negligible and not significant. The relationships between both GMA ($r = .24, p < .05$) and agreeableness ($r = .18, p < .05$) and should-SJT scores were moderate. Furthermore and similarly, z-tests involving Fisher’s r-to-z transformations indicate that both GMA/SJT correlations were statistically different ($p < .05$), while both agreeableness/SJT correlations were not ($p > .05$). Taking these results together, Hypothesis 3 was not supported through the correlational analyses.

As a more statistically sophisticated complement to the initial set of correlational analyses, a moderated multiple regression was also run examining the same issue. That is, a multiple regression model tested whether the association between constructs of interest and SJT scores depends on the instruction set (i.e., instruction set is the

Table 7. Hypothesis 3: Correlations between Would-do and Should-do SJT scores and GMA and Agreeableness (excluding participants who failed post-5th item MC)

	Would SJT scores	Should SJT scores
GMA	↓ -.03 ^a (<i>n</i> = 112)	↑ .24* ^a (<i>n</i> = 87)
Agreeableness	↑ .04 ^b (<i>n</i> = 112)	↓ .18* ^b (<i>n</i> = 87)

Note. ^aCorrelations are significantly different ($p < .05$) ^bCorrelations are not significantly different ($p > .05$). * $p < .05$, one-tailed. The vertical arrows represent the posited effects such that ↑ represents an expected stronger relationship and ↓ a weaker one.

moderator). After converting the GMA scores, agreeableness scores, and SJT scores to z-scores, the construct of interest (i.e., GMA or agreeableness), instruction set, and the interaction term were entered into a regression model as predictors in a hierarchical fashion. Standardizing variables in the regression model is done for interpretability reasons. Doing so simplifies making comparisons across models and the newly standardized beta coefficient reflects the correlation between predictor and outcome.

If the instruction set moderates the relationship between the constructs of interest and SJT scores, then the interaction term should be significant, thus indicating the interaction term contributes to explaining variance in the criterion. Results indicated that instruction set did not moderate the relationship between GMA scores and SJT scores ($b = .23, SE = .12, \beta = .16, p > .05$) or the relationship between agreeableness scores and SJT scores ($b = .16, SE = .12, \beta = .12, p > .05$). This indicates that for the relationship between GMA and SJT scores *and* the relationship between agreeableness and SJT scores, the SJT instruction set was not a moderator. That is, neither of the interactions for Hypothesis 3 (i.e., for GMA or agreeableness) were significant. Therefore, in

addition to lacking support from the correlational analyses, Hypothesis 3 was not supported by the regression analyses as well.

As a follow-up test to this analysis, the same regressions were conducted excluding the participants who failed the post-5th item MC, due to the same logic presented previously. Again, although the ideal sample would *only* include participants who recalled the focal word at item 1, the current analyses was conducted excluding those participants who did not recall the word at item 5 (indicating for these people that the focal word was not processed even by the 5th item). Excluding these participants did not alter the results. The instruction set did not moderate the relationship between agreeableness and SJT scores ($b = 2.39, SE = 2.68, \beta = .09, p > .05$) or the relationship between GMA and SJT scores ($b = 5.35, SE = 2.87, \beta = .18, p > .05$). Therefore, this further demonstrates lack of support for Hypothesis 3.

Research Question 1 stated that *“If participants recall the focal word in the MC item(s), then is the pattern of results found in the present study construct-invariant such that the relationships found between SJT scores (with would-do and should-do instruction sets) and GMA/personality traits (i.e., agreeableness) remain unchanged when different constructs are assessed (cognitive vs. noncognitive constructs)?”* To examine this question, SJT scores for the conditions (i.e., standard and/or bold/underline) in which the participants noticed the focal word are of main interest. However, similar to Hypothesis 3, these analyses were conducted twice, once with the whole sample, and once again excluding participants who did not pass the post-5th MC

item, as a supplemental exploratory analysis. Both correlations and regression models were used to examine this research question as well.

Regarding the main analyses for Research Question 1, only participants who passed the post-5th item MC were considered. This is considered a conservative test of this research question (i.e., one that reduces Type 1 error rate) because it is based on the inclusion of participants who conceivably could have failed to recall all MC's preceding the last MC. However, due to practical constraints, this boundary condition was chosen. Four sets of pairwise correlations comprise the first prong of the analytic strategy for this research question. These correlations compare the relationships between constructs of interest (i.e., GMA and agreeableness) and SJT scores by instruction set and construct assessed. Results are provided in Table 8. In summary, the observed pattern of relationships displayed here does not fit any discernable pattern. Therefore, the construct assessed was not shown to influence the effect of the instruction set on the construct-related validity of SJTs in any particularly patterned way in terms of the posited effects.

The potential effect of the construct assessed was further examined using regression as well. Specifically, moderated multiple regression was used to test whether the effect of the instruction set was moderated by the construct (GMA and agreeableness) assessed by the SJT, and the results indicated that the relationship between GMA and would-SJT/should-SJT scores was not influenced by the construct assessed ($b = -.03$, $SE = .28$, $\beta = -.01$, $p > .05$). Similar results were obtained for

Table 8. Research Question 1: Correlations between Would-do and Should-do SJT scores and GMA and Agreeableness (excluding participants who failed post-5th item MC) by Construct

	Would SJT scores		Should SJT scores	
	Cognitive		Cognitive	
	Agreeableness		Agreeableness	
GMA	↑ .04 (<i>n</i> = 53)	↓ .02 (<i>n</i> = 59) ↓	↑ .24 (<i>n</i> = 39)	↑ .12 (<i>n</i> = 48) ↓
Agreeableness	↓ -.16 (<i>n</i> = 53)	↑ .20 (<i>n</i> = 59) ↑	↓ .24 (<i>n</i> = 39)	↓ .28* (<i>n</i> = 48) ↑

Note. The vertical arrows represent the posited effects such that up arrows (↑/↑) represent an expected stronger relationship and down arrows (↓/↓) a weaker one. Furthermore, thick arrows (↑↓) represent construct of interest/instruction set posited relationships (as in Table 6), while thin arrows (↑/↓) reflect further posited relative strength of relationships with construct assessed. Relationships with two up arrows on either side represent the combination of variables with the strongest posited relationships. Only one out of these four pairs of correlations (same construct, between column comparisons) were significantly different to each other. Specifically, for the cognitive SJT, agreeableness relationships with SJT scores were statistically different between should and would versions ($p < .05$). * $p < .05$, one-tailed.

agreeableness ($b = -.25$, $SE = .26$, $\beta = -.14$, $p > .05$). This further confirms that the construct assessed by the SJT failed to interact with the instruction set.

To further examine this question, SJT scores within all conditions, regardless of whether participants accurately recalled the focal word, were analyzed. The first prong of this supplemental analyses is, again, examining pairwise correlations. Results are provided in Table 9. Similar to the previous analysis, it is worth noting that the array of observed relationships displayed here does not follow the pattern predicted by the posited effects.

Again, the second prong of this analysis involved the use of moderated multiple regression to examine whether the effect of the instruction set was moderated by the construct assessed by the SJT. Results indicated that for the relationship between GMA

Table 9. Research Question 1: Correlations between Would-do and Should-do SJT scores and GMA and Agreeableness by Construct

	Would SJT scores			Should SJT scores		
	Cognitive	Agreeableness		Cognitive	Agreeableness	
GMA	↑ .09 (<i>n</i> = 63)	↓ -.02 (<i>n</i> = 65)	↓	↑ .31* (<i>n</i> = 66)	↑ .10 (<i>n</i> = 70)	↓
Agreeableness	↓ -.06 (<i>n</i> = 63)	↑ .19 (<i>n</i> = 65)	↑	↓ .23* (<i>n</i> = 66)	↓ .32* (<i>n</i> = 70)	↑

Note. The vertical arrows represent the posited effects such that up arrows (↑/↑) represent an expected stronger relationship and down arrows (↓/↓) a weaker one. Furthermore, thick arrows (↑↓) represent construct of interest/instruction set posited relationships (as in Table 6), while thin arrows (↑/↓) reflect further posited relative strength of relationships with construct assessed. Relationships with two up arrows on either side represent the combination of variables with the strongest posited relationships. Out of these four pairs of correlations (same construct, between column comparisons), none were significantly different from each other. **p* < .05, one-tailed.

and SJT scores, the construct assessed was not a significant moderator ($b = -.01$, $SE = .22$, $\beta = .00$, $p > .05$). Furthermore, similar results were obtained for agreeableness ($b = -.11$, $SE = .21$, $\beta = -.06$, $p > .05$). This further confirms the lack of support for any discernable influence the construct assessed by the SJT has on the instruction set effect.

Hypothesis 4 stated that “*Cognitive SJTs with should-do instruction sets will be more highly correlated with GMA than cognitive SJTs with would-do instruction sets,*” and Hypothesis 5 stated that “*Noncognitive SJTs with would-do instruction sets will be more highly correlated with personality traits (i.e., agreeableness) than noncognitive SJTs with should-do instruction sets*”. Testing of these hypotheses was predicated on the assumption that the instruction set effect garnered support. Because the instruction set effect was not supported in the present study, these hypotheses are not formally presented here. However, for the sake of completeness, the results regarding these hypotheses are reported in Appendix B.

DISCUSSION AND SUMMARY

A number of summary statements can be made about the results of the present study. First, respondents generally did not recall the focal word (i.e., would or should) embedded in the instructions of the SJTs, although there appeared to be a priming and repetition effect. Second, for the most part, a **bold/underline** modification of the focal word saliency did not substantially increase recall accuracy, operationalized as a response to a MC item asking the respondent to fill in the focal word. Third, the instruction set effect was not replicated in the present study, which used a single word manipulation with at least one MC (which varied by condition). Fourth, the construct assessed by the SJT did not affect the instruction set effect or lack thereof.

To further expand on these summary statements, first, because first-encounter MC pass rates were generally low (i.e., pass rates were no different from a 50% base rate), it can be presumed that test takers in applied settings would generally fail to accurately recall the focal word in standard test-taking conditions if given an MC. That is, because typical SJTs present the focal word in the instruction set in a standard way (i.e., with no saliency modifications), the finding that participants do not recall the focal word in the standard conditions suggests that test-takers in applied settings likely do not notice the focal word either.

It is not surprising that participants did not recall the focal word of the SJTs in the standard conditions (as predicted in Hypothesis 1). Based on the tenets of the literature on anticipation and eye movements during word processing, the focal word seems to be a word that can be easily glossed over. Specifically, because anticipation is

likely to decrease depth of processing and readers are likely to skim over helping verbs like “would” and “should”, it was not unreasonable to expect that participants would not recall the instruction set focal word. Instead, a more likely explanation is that the focal word was glossed over and not processed because its meaning and function in the sentence could be anticipated.

Second, there are conceivably a number of ways to increase test takers' attention to the focal response instruction word; the present study examined two of these, the first of which was to increase the saliency of the focal word by bolding and underlining it. It *is* surprising that the **bold/underline** modification did not result in a substantial increase in the recall pass rates. Because of the results of informal pilot tests which showed a dramatic increase in pass rates when students in classrooms were presented with focal words that were bolded or underlined, it was expected that pass rates in the saliency conditions would be much higher (as predicted in Hypothesis 2) than they were. Although there were some small increases in the hypothesized direction, they were generally not significantly different from those in the standard conditions. It is nevertheless, noteworthy that for the first and third MC items, there was a noticeable increase (about 17%; although not statistically significant) in pass rates between the standard and **bold/underline** conditions (MC 1, 36% to 54%; MC 3, 46% to 63%). Therefore, if one seeks to increase short-term, initial processing of the focal word (i.e., during the first few items of an SJT), the **bold/underline** modification might be useful. However, by the fifth item, there was no noticeable difference in pass rates between the standard and **bold/underline** conditions (51% and 52%, respectively). Because the pass

rates converged by the post-5th item MC, the **bold/underline** modification only had short-term gains in facilitating recall, and by inference, processing. Furthermore, and by extension, given that this manipulation was a combination of two saliency-inducing modifications, **bold** and underline, these results suggest that using either modification in isolation will likely have limited utility.

One potential explanation for this levelling off by the 5th item is that the repetition effect supersedes the saliency effect by this point. That is, after having seen 5 items that contain the instruction set, the effect of the saliency modification on the pass rate becomes unnoticeable due to the number of times participants have seen the response instructions. This repetition effect can be considered to be another way to increase the accurate recall of the focal word (i.e., through repeated exposure).

Although the **bold/underline** intervention turned out not to be particularly effective, another intervention was. Specifically, asking participants to recall the focal word after an SJT item noticeably and significantly increased recall pass rates assessed after this item. That is, asking participants to recall the focal word using a fill-in-the-blank response seemed to prime participants to pay attention to the focal word in subsequent items. While not formally presented as an intervention in the present study, the first round of MCs (i.e., in conditions where participants were given two MCs) permitted an examination of a priming effect and thus can be seen as an intervention (i.e., a prime) to increase processing. The second MC in the present study can still be regarded as the traditional MC, as this is always given after the last SJT item in instances when there are two MCs and still accomplishes the goal of confirming whether the focal

word was recalled. This priming strategy was more effective at increasing accurate recall of the focal word (in Figure 4, see the comparison between 5B [first-encounter post-5th item pass rate] and 5C [second-encounter post-5th item pass rate]), and generally resulted in pass rates around 80-90%.

In Table 2, comparing 5B (post-5th item MC only condition) to 5C (post-5th item second-encounter only), pass rates increased from 51% to 84% for the standard condition, and 52% to 90% for **bold/underline** conditions. Consequently, if one has an interest in having the focal word noticed and processed as early in an SJT as possible, then administering a recall MC after the 1st item appears to be a promising way of ensuring that test takers will pay close attention to the focal word in subsequent items.

Third, the null findings regarding the instruction set effect calls the purported received doctrine (i.e., that should-SJTs behave like ability measures and would-SJTs behave like personality measures) into question. It was not surprising that the instruction set effect was not replicated (and Hypothesis 3 was thus not supported) because the extant body of literature on which this effect is based contains notable methodological concerns. Specifically, past research has confounded changes to response instructions with changes to other parts of the SJT item stem (e.g., response format; e.g., McDaniel et al., 2007, Ployhart & Ehrhart, 2003; and other parts of the stem; see Figure 1 for a list of examples). The one study that employed a single word manipulation (Kelly, 2013) also obtained null results but did not include a MC. Therefore, because previous research has included manipulations with multiple changes to the SJT items and resulted in obtaining instruction set effects, it is unclear which

particular change led to these effects. Furthermore, past research has not consistently used a MC; therefore, it is unclear if the focal words were processed in these instances. The present study addressed these concerns by employing a single word manipulation with a MC and did not replicate this effect.

Consequently, to the best of our knowledge, the present study represents the most rigorous test of the instruction set effect to date. So, given that the present study failed to obtain the instruction set effect with its specified methodological characteristics, this serves as the basis for a call for further research to investigate why previous studies examining the instruction set effect found support for it. This issue is further discussed as a direction for future research in a subsequent section.

Fourth, although the instruction set effect was not replicated and Hypothesis 3 was not supported, the theoretical underpinnings of the instruction set effect are still quite conceptually compelling: that would-do instructions and personality tests both engender a behavioral-based mindset, while should-do instructions and cognitive ability tests both engender a knowledge-based mindset. Thus, additional examinations of this notion are warranted. However, it is ultimately unlikely that a single word that is glossed over is responsible for changing the psychological processes engendered by the instructions to any notable extent. Therefore, it seems that the instruction set effect, as currently purported, may be a received doctrine awaiting compelling empirical support.

Fifth, because SJTs are a method, and a research question of the present study was to examine the instruction set effect in the context of the constructs the SJTs were designed to assess, the construct assessed by the SJT was examined to investigate

whether it might influence the instruction set effect or lack thereof. Not only is it a general best practice to distinguish between constructs and methods and conduct research on methods in the context of the constructs they are designed to assess (Arthur & Villado, 2008), but it is also conceptually conceivable that the construct assessed might further influence the instruction set effect. Therefore, it was reasonable to examine whether the constructs of interest and instruction sets might interact with the construct assessed; however, the direction of the effect, if any, was unknown. Thus this issue was posed as a research question because it was exploratory in nature.

The results indicated that the construct assessed does *not* interact with the instruction set to influence the construct-related validity of SJTs. Specifically, in light of the instruction set not having a direct influence on the construct-related validity of SJTs, the construct assessed does not interact with the instruction set effect to influence the construct-related validity of SJTs. Therefore, the lack of support for the instruction set effect appears to be construct-invariant.

It is worth noting that the construct assessed *does* influence the construct-related validity of SJTs as one might predict (in terms of a main effect), and this is evidenced by the expected pattern of correlations with different constructs of interest (as reported in the Results). That is, the cognitive SJT is more highly related to GMA than agreeableness, and the agreeableness SJT is more highly related to agreeableness than GMA. In light of this, the finding that the instruction set effect is not further influenced by the construct assessed is consonant with the finding that the instruction set effect was not supported.

If the instruction set effect had been supported, then matching the construct assessed and construct of interest with the aligned instruction set (i.e., thought to engender the same psychological processes) would likely reduce construct-irrelevant variance and increase the strength of the relationships containing all of the aligned characteristics (Hypothesis 4 and 5). However, because the instruction set was not supported, and Hypothesis 4 and Hypothesis 5 were based on garnered support for the instruction set effect, the hypothesized effects regarding alignment between these characteristics was not found (see Appendix B for these results).

Implications

Given the prevalence of SJTs in applied settings (likely due to their strong criterion-related validities and potentially lower levels of subgroup differences), studying factors that influence their construct-related validity is a worthwhile endeavor. The extant literature purports the instruction set effect, but has been based on studies that have had methodological concerns. Consequently, the failure to replicate this effect using a single word manipulation (with a MC) has notable implications.

Implications for science. First, contrary to the widely-held doctrine that response instructions influence the construct-related validity of SJTs, the present study suggests that the focal word does not impact this as advanced in the literature. The implications of these null findings suggest that the instruction set effect has been received as a doctrine in the absence of compelling evidence. Therefore, this influences how instruction sets should be viewed as a design feature and limits the importance that can be placed on their influence on the construct-related validity of SJTs. Specifically, it

appears that should-SJTs do not act like ability tests, and would-SJTs do not act like behavioral measures. Furthermore and similarly, should-SJTs do not engender a knowledge-based, cognitive orientation, while would-SJTs do not engender a behavioral orientation. Second, if previous findings are simply due to confounds or methodological weaknesses, then it begs the question about what other factors (or combination of factors) or features of previous studies could have caused these findings (e.g., response format; see Figure 1).

Implications for practice. There are also practical implications regarding what one can do to increase processing of the focal word in the instruction set of SJTs (or potentially any word or phrase, for that matter). One effective way to increase processing of the focal word seems to be by providing a prime (e.g., a question asking for a recall of the word). This seems to increase subsequent processing of the word because all the post-fifth item, second-encounter MCs were higher than their first-encounter counterparts (see Figure 4). Therefore, providing a prime post item 1 or post item 3 substantially increased the recall pass rate at the MC post item 5. Hence, to the extent that the processing of the focal word—would/should—is of importance to the researcher or practitioner, then inserting a recall prime after the first item, will prompt the test taker to pay attention to the word on subsequent items. While processing the focal word does not seem to matter regarding SJT response instructions, it may still be germane to processing words of interest to certain SJTs or more broadly, other employment-related tests. In cases where processing of the focal word is critical to the purpose of the test, using this strategy also implies that the post-priming item might

more appropriately be considered the test's first item (while the pre-priming item might be more of a sample or practice item).

There is also an implication regarding the repetition effect present in focal word processing. The present study provides preliminary evidence on *when* participants notice the instruction's focal word under typical testing circumstances (when no saliency modifications or priming mechanisms are present). The standard conditions show a positive trend between MC pass rates and the number of items that have been completed before the first MC (i.e., pass rates of interest are 36%, 46%, and 51%, for MC items 1, 3, and 5, respectively). Therefore, this demonstrates preliminary evidence for a repetition effect such that the more items that a participant is exposed to, the more likely the focal word will be processed. On the other hand, the **bold/underline** conditions displayed a more nuanced (and less clean) trend in this regard. Although there seems to be an increase (between-subjects) in pass rates from MC1 to MC3, there is then a slight decrease between MC3 and MC5 across constructs (i.e., pass rates of interest are 54%, 63%, and 52%, for MC items 1, 3, and 5, respectively). There is no compelling conceptual basis for this slight decrease, and thus any attempt to explain it would at best be speculative.

Another implication pertains to the design of SJTs. Whereas the literature might lead one to conclude that the choice of response instructions influences the construct-related validity of SJTs such that should-SJTs engender more of a knowledge-based orientation, while would-SJTs engender more of a behavioral orientation, the results of the present study in which only the focal word, would versus should, was manipulated

calls this received doctrine into question. Therefore, in designing SJTs, the implications resulting from the decision between instruction sets are limited.

The final implication concerns the extent to which the present study can potentially provide guidance on the type of instruction set to use for SJTs with different content. If Hypotheses 4 and 5 had been supported, then this would have suggested that one should use should-do instruction sets for cognitively-loaded SJTs and would-do instruction-sets for noncognitively loaded SJTs. Therefore, additional practical implications would have potentially included the identification of the instruction set that would be most appropriate to reduce construct-irrelevant variance, given the construct measured. However, because these hypotheses were not supported, there is no conclusive guidance regarding which instruction sets to use in different cases.

Limitations, Directions for Future Research, and Conclusions

Limitations. There are limitations of present work that need to be acknowledged. First, the power levels of the study are relatively low. Indeed, given the observed effects, a sample size of 834 would be required to achieve power of .80, and 2,896 to achieve a power of .99. However, although it will likely alter its significance, it is unlikely that additional data will reverse the magnitude and/or pattern of results reported here. Second, this was a low-stakes assessment using college students. However, it is unclear that high-stakes job applicants would necessarily obtain higher recall pass rates. Indeed it could be reasonably argued that job applicants are likely to approach the SJT as a "should" assessment regardless of the response instructions. That is, in high-stakes assessments, applicants are likely to report what they consider to be the maximally

correct response instead of what *they* would typically do. This is noteworthy since the preponderance of the SJT research has used incumbents (Ployhart & MacKenzie, 2011).

Third, a potential explanation for null results could be that would-SJTs become should-SJTs in high-stakes situations, such as the life-threatening survival scenarios presented in the cognitive SJT used in the present study. In contrast to the previous limitation, one could potentially argue that the survival context has a high-stakes component. While the test-taking situation was not high-stakes according to external contingencies (i.e., as in a real job application), it was high-stakes according to hypothesized contingencies (i.e., as portrayed by the life-threatening survival situations). The logical extension of this notion is that in high-stakes settings, would-do and should-do instructions become indistinguishable. In these settings, both response instructions function as should-do instructions because test-takers would have an inherent incentive to perform well by providing the correct, knowledge-based answer (as opposed to their behavioral answer). Therefore, if this is a limitation, then using an SJT with a more normal context (i.e., non-life-threatening) might be an option to consider in future research that seeks to further examine the issues addressed here. However, this does not represent a major limitation of the present study, as the instruction set effect proponents would argue the effects on the construct-related validity would still be present even in SJTs employing a survival context.

To the extent that the nature of the cognitive SJT is a potential limitation to the present study, it is important to note that *only* the results of the cognitive SJT would be placed into question. Therefore, although this critique might decrease confidence in the

pattern of results obtained for the cognitive SJT, the agreeableness SJT results are unaffected by this limitation. Consequently, focusing exclusively on the agreeableness SJT, the results presented in Table 8 indicate that the pattern displayed by the agreeableness columns still cannot be said to be in alignment with the purported instruction set effect (e.g., the agreeableness/should relationship with agreeableness is actually *higher* than the agreeableness/would relationship with GMA). In summary, the potential limitation posed by the atypical nature of the cognitive SJT only relates to a subset of the results to begin with, and therefore, does not alter the conclusions of the present study.

Directions for future research. Additional research is encouraged to examine why previous studies might have found this notable, yet confounded shift in construct-related validity, which has been previously attributed to response instructions (see Figure 1). This study's results call to question the findings of previous studies that were conducted using manipulations involving more than a single word change. For example, while Ployhart and Ehrhart (2003) found that different types of response instructions influence SJT scores, reliability, criterion-related validity (with a variety of criteria) and construct-related validity, these findings were observed while using a confounded manipulation. Therefore, further examination of these issues is warranted using a single word manipulation with a manipulation check.

With that as a backdrop, the present study *can* tentatively speak to a subset of these additional issues—mean scores, variability, and reliability. The present study obtained an internal consistency reliability estimate of .86 for both the cognitive *should-*

do and *would-do* SJTs. Similarly, the present study also obtained an internal consistency reliability estimate of .91 for the agreeableness *should-do* and *would-do* SJTs. Thus in both instances, the internal consistency reliability estimates did not change as a function of the instruction set. Furthermore, the difference between the means for cognitive *would-do* and *should-do* SJT scores was negligible ($d = 0.06$, $M = 32.06$ and 32.80 , respectively); with an equally small difference in the standard deviations ($SD = 12.75$ and 11.44 for *would-do* and *should-do*, respectively). Similar results were obtained for the agreeableness SJT. Again the mean difference in *would-do* and *should-do* SJT scores was negligible ($d = 0.07$, $M = 50.92$ and 52.29 , respectively), with an equally small difference in the standard deviations ($SD = 19.26$ and 18.22 for *would-do* and *should-do*, respectively).

The tentative implications of these preliminary results are in accordance with the major findings of the present study. Namely, the choice of instruction set has less of an impact on these outcomes, compared to what is currently advanced in the SJT response instructions literature. Therefore, because the present study reported null findings with a single word manipulation, in contrast to previous research that has been conducted using manipulations involving more than just a single word change, all of these other issues reviewed (e.g., criterion-related validity) should be revisited in the context of a single word manipulation. Further examination of the impact of the instruction set on faking is also warranted, including the findings by Nguyen, et al. (2005) that indicate *would-do* SJTs are easier to fake than *should-do* SJTs. Lievens, Sackett, and Buyse (2009) did not replicate this fakeability finding in a high-stakes setting. However, because neither

study employed a single word manipulation, it is unknown whether would-SJTs and should-SJTs in both low stakes and high stakes settings are differentially susceptible to faking.

Furthermore, research that further explores when participants process words focal to the purpose of the SJT, such as instruction set focal words, is also called for. At a broader level, it might also be informative to conduct further research regarding the timing of focal word processing of other predictor methods that contain focal words or phrases. Finally, research examining the conceptual underpinnings of the instruction set effect, regarding how to elicit a knowledge-based vs. behavioral-based orientation would be informative and is thus warranted.

Conclusion. In conclusion, the goal of the present study was to re-examine the SJT response instruction effect. Because this effect is (1) widely reported in the personnel testing and assessment literature but is (2) based on a foundation of studies with some methodological concerns, it was re-examined using a clear, succinct single-word manipulation. An attempt to replicate the instruction set effect failed to provide support for what can be described as a received doctrine, that should-do instructions result in high correlations with ability and low correlations with personality traits, and would-do instructions in contrast, result in high correlations with personality but low correlations with ability. Furthermore, this failure to replicate the instruction set effect is observed for both a cognitive and noncognitive SJT. Finally, the construct assessed had no influence on the instruction set effect, or lack thereof. Future research using large samples in high-stakes assessments is called for.

REFERENCES

- Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology, 90*, 94-107.
- Ang, S., Van Dyne, L., Koh, C., Ng, K. Y., Templer, K. J., Tay, C., & Chandrasekar, N. A. (2007). Cultural intelligence: Its measurement and effects on cultural judgment and decision making, cultural adaptation and task performance. *Management and Organization Review, 3*, 335-371.
- Arthur, W., Jr. (2017a). Construct-laden situational judgment tests of personality traits: Ingenuity or folly? In J. Golabovich (Chair), *Development and scoring of construct-focused situational judgment tests*. Symposium to be presented at the 32nd annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.
- Arthur, W., Jr. (2017b). *An unproctored Internet-based test of general mental ability. Validation report*. College Station, TX: Author.
- Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435-442.
- Arthur, W., Jr., Glaze, R. M., Jarrett, S. M., White, C. D., Schurig, I., & Taylor, J. E. (2014). Comparative evaluation of three situational judgment test response formats in terms of construct-related validity, subgroup differences, and susceptibility to response distortion. *Journal of Applied Psychology, 99*, 535-545.
- Bauer, T. N., & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekley, & R. E. Ployhart (Eds.),

- Situational judgment tests: Theory, measurement, and application* (pp. 223-249).
Mahwah, NJ: Lawrence Erlbaum Associates.
- 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.
- Braun, K. A., & Loftus, E. F. (1998). Advertising's misinformation effect. *Applied Cognitive Psychology, 12*, 569-591.
- Busch, J. C., & Jaeger, R. M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*, 145-163.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Campion, M. C., Ployhart, R. E., & MacKenzie Jr, W. I. (2014). The state of research on situational judgment tests: A content analysis and directions for future research. *Human Performance, 27*, 283-310.
- Carpenter, P. A., & Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes* (pp. 275-307). New York: Academic Press.
- 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.

- Christian, M. S., Edwards, B. D., and Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83-117.
- 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.
- Cunitz, R. J., & Steinman, R. M. (1969). Comparison of saccadic eye movements during fixation and reading. *Vision Research, 9*, 683-693.
- Day, E. A., Arthur, W., Jr., Miyashiro, B., Edwards, B. D., Tubré, T. C., & Tubré, A. H. (2004). Criterion-related validity of statistical operationalizations of group general cognitive ability as a function of task type: Comparing the mean, maximum, and minimum. *Journal of Applied Social Psychology, 34*, 1521-1549.
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28*, 1-11.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.
- Goodman, K. S. (1979). The know-more and the know-nothing movements in reading: A personal response. *Language Arts, 56*, 657-663.

- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23, 17-27.
- Holmes, V. M., Arwas, R., & Garrett, M. F. (1977). Prior context and the perception of lexically ambiguous sentences. *Memory & Cognition*, 5, 103-110.
- Johnson, D. W., & Johnson, F. P. (1991). *Joining together: Group theory and group skills* (4th ed.). Englewoods, NJ: Prentice-Hall, Inc.
- Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view. *European Journal of Psychological Assessment*, 22, 168-176.
- Kelly, C. D. (2013). *The effects of response instructions on situational judgment test performance in high-stakes employment context* (Master's thesis). California State University, Sacramento.
- Kolers, P. A. (1970). Three stages of reading. In H. Levin & J. Williams (Eds.), *Basic studies on reading* (pp. 90-118). New York, NY: Basic Books.
- Knox, G. (2009). *Lost at sea* [technical report]. Retrieved from http://insight.typepad.co.uk/lost_at_sea.pdf
- Lafferty, J. C. (1987). *Jungle survival situation: Leader's guide*. Plymouth, MI: Human Synergistics.
- LePine, J. A., Colquitt, J. A., & Erez, A. (2000). Adaptability to changing task contexts: Effects of general cognitive ability, conscientiousness, and openness to experience. *Personnel Psychology*, 53, 563-593.
- Lievens, F., De Corte, W., & Westerveld, L. (2015). Understanding the building blocks of selection procedures effects of response fidelity on performance and validity. *Journal of Management*, 41, 1604-1627.

- Lievens, F., & Patterson, F. (2011). The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *Journal of Applied Psychology, 96*, 927-940.
- Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*, 426-441.
- Lievens, F., Sackett, P. R., & Buyse, T. (2009). The effects of response instructions on situational judgment test performance and validity in a high-stakes context. *Journal of Applied Psychology, 94*, 1095-1101.
- Loftus, E. F. (1977). Shifting human color memory. *Memory & Cognition, 5*, 696-699.
- Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589.
- Loftus, E. F., & Zanni, G. (1975). Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society, 5*, 86-88.
- MacKenzie, W. I., Jr., Ployhart, R. E., Weekley, J. A., & Ehlers, C. (2009). Contextual effects on SJT responses: An examination of construct validity and mean differences across applicant and incumbent contexts. *Human Performance, 23*, 1-21.
- 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.
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. (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., Psotka, J., Legree, P. J., Yost, A. P., & Weekley, J. A. (2011). Toward an understanding of situational judgment item validity and group differences. *Journal of Applied Psychology, 96*, 327-336.
- 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., Ghosh, K., Mendoza, A. M., Buchanan, A. E., & Lerma, M. N. (2016). A context-independent situational judgment test to measure prosocial implicit trait policy. *Human Performance, 29*, 331-346.
- Naber, A. M., Arthur, W., Jr., Edwards, B. D., & Franco-Watkins, A. (2017). *Increased retest scores on cognitive tests: Learning or memory effects*. Unpublished manuscript.
- Nguyen, N. T., & McDaniel, M. A. (2003). Response instructions and racial differences in a situational judgment test. *Applied HRM Research, 8*, 33-44.
- 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.
- Olson, J. N., & MacKay, D. G. (1974). Completion and verification of ambiguous sentences. *Journal of Verbal Learning and Verbal Behavior, 13*, 457-470.

- Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51*, 1-24.
- Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management, 32*, 868-897.
- Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153-172.
- 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.
- Ployhart, R. E., & MacKenzie, W. I., Jr. (2011). Situational judgment tests: A critical review and agenda for the future. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, Vol. 2: Selecting and developing members for the organization* (pp. 237-252). Washington D.C., U.S.: American Psychological Association.
- Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition, 5*, 443-448.
- Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85*, 618-660.
- Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880-887.

- Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*, 297-330.
- Schmauder, A. R., Morris, R. K., & Poynor, D. V. (2000). Lexical processing and text integration of function and content words: Evidence from priming and eye fixations. *Memory & Cognition, 28*, 1098-1108.
- Shebilske, W. L. (1975). Reading eye movements from an information-processing point of view. In D. Massaro (Ed.), *Understanding language: An information-processing analysis of speech perception, reading, and psycho-linguistics* (pp. 291-311). New York, NY: Academic Press.
- Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence, 53*, 92-101.
- Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 443-467.
- Wanat, S. F. (1976). Relations between language and visual processing. In H. Singer & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (pp. 108-136). Newark, DE: International Reading Association.
- Weekley, J. A., Hawkes, B., Guenole, N., & Ployhart, R. E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 295-322.

- 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.
- Weinstein, C.E. (1987). *LASSI User's Manual*. Clearwater, FL: H & H Publishing Company, Inc.
- Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, *21*, 291-309.
- Wildman, D., & Kling, M. (1978). Semantic, syntactic, and spatial anticipation in reading. *Reading Research Quarterly*, *14*, 128-164.

APPENDIX A

SITUATIONAL JUDGMENT TEST (SJT) SAMPLE ITEMS

Cognitive SJT

This measure consists of 5 survival scenarios in different settings. Each scenario presents a situation in which you are stranded in a harsh environment and have salvaged **four items** that can be used to help you survive. Your task is to rank the 4 salvaged items in terms of their importance for your survival.

Please rank order the five items in terms of their importance for your survival. Note that **1 = MOST important for survival** and **4 = LEAST important for survival**.

Please enter your rank in each space provided. **There cannot be any ties** in the ranks.

At the end of each page, use the right arrow button on the bottom right corner to move to the next page. **Do not use your browser's back button while completing this activity.**

1. You and one companion have just survived the crash of a small plane somewhere in the Brazilian rain forest. Both the pilot and co-pilot were killed. You are surrounded by wet, thorny foliage, and you smell gasoline coming from the plane. As the sun is setting, you manage to salvage 4 items that could be used to help you survive these conditions. Rank the following items in terms of how important you would consider them to be in terms of what you need to do to survive until you are rescued.

		Rank	Keyed	Scenario
1.	Plastic "space blanket" (7' × 5.5"); a multi-purpose tool (e.g., drying kindling during the day to start the night's fire, carrying items during day travels, providing shelter and fire at night by rigging it above)		2	4
2.	1 box of granola bars; for food energy		3	8
3.	.38-caliber pistol (loaded); sound-signaling device and weapon		4	12
4.	Machete; all-purpose tool (e.g., cutting vines containing drinking water, cutting branches/palms to build shelter, cutting paths through foliage)		1	1

Agreeableness (noncognitive) SJT

This measure describes 5 hypothetical workplace scenarios. For each scenario, please rank order the actions in terms of their effectiveness as a solution or response to the situation presented in the scenario. That is, your task is to rank the four actions in terms of their effectiveness.

Please rank order the four actions in terms of their effectiveness as a solution to the situation.

Note that **1 = MOST effective** and **4 = LEAST effective**. There **cannot be any ties** in the ranks.

1. You and your coworker, Alex, have a meeting scheduled to take care of the next steps on a project. However, Alex has rescheduled the meeting twice in the last week. You have both agreed to meet today to discuss the project but Alex has just rescheduled the meeting again. Rank the effectiveness of the following in terms of what you would do.

		Rank
A.	Send Alex an email letting her know that the project needs to be completed in a timely manner.	
B.	Ask Alex for a new meeting date that works for her.	
C.	Send Alex an email letting her know how unprofessional this behavior is.	
D.	Ask to speak to Alex for a few moments with the intent of professionally reminding her about the importance of keeping commitments and respecting each other's time.	

APPENDIX B

RESULTS OF HYPOTHESES 4 AND 5

Hypothesis 4 stated that “*Cognitive SJTs with should-do instruction sets will be more highly correlated with GMA than cognitive SJTs with would-do instruction sets,*” and Hypothesis 5 stated that “*Noncognitive SJTs with would-do instruction sets will be more highly correlated with personality traits (i.e., agreeableness) than noncognitive SJTs with should-do instruction sets.*” To test these hypotheses, both correlations and moderated multiple regressions were conducted between GMA/agreeableness and SJT scores measuring particular constructs, with particular instruction sets as follows. Like Research Question 1, moderated multiple regression models were tested to examine whether the association between constructs of interest and SJT scores depends on how the construct assessed and instruction set interact with each other. Because the regression analyses for Research Question 1 showed that the construct assessed variable did not moderate the relationship between construct of interest (i.e., GMA or agreeableness) and would-SJT/should-SJT in case of either construct of interest, this section focuses on the trends that can be observed from the correlations. Please see Table B1 for a summary of the relationships.

Regarding Hypothesis 4, the correlation between GMA and cognitive SJT scores with the *should* instruction set was moderate-to-large ($r = .31, p < .05$), while the correlation between GMA and cognitive SJT scores with the *would* instruction set was considerably lower ($r = .09, p > .05$). A z -test using Fisher’s r -to- z transformation indicated these two correlations were not significantly different ($p > .05$). Regarding Hypothesis 5, the correlation between agreeableness and agreeableness SJT scores with the *should* instruction set was moderate ($r = .32, p < .01$), while the correlation between agreeableness and cognitive SJT scores with the *would* instruction

set was lower ($r = .19, p > .05$). A z -test using Fisher's r -to- z transformation indicated these two correlations were not significantly different ($p > .05$). Overall, the construct assessed does not seem to further influence the instruction set effect in any predictable way. Essentially, these hypotheses consisted of the predictions that when the instruction set and the construct assessed is aligned (which reduces construct-irrelevant variance), the correlations between the construct of interest and SJT scores would be highest. This set of predictions was not supported.

Furthermore, it is important to note that one out of two correlations within the trend is in the opposite direction than expected (i.e., the relationship between agreeableness and SJT scores is actually higher when the construct assessed and instruction set are *not* aligned with agreeableness). Therefore, although it seems like the initial correlational trend (between GMA and *cognitive* should-SJT vs. GMA and *agreeableness* should-SJT) is in line with Hypothesis 4, the moderated multiple regression still displays lack of support for any strong or consistent influence on the instruction set effect by the construct assessed. The construct assessed cannot be said to interact with the instruction set in any patterned way. Therefore, Hypotheses 4 and 5 were not supported.

Table B1. Hypotheses 4 and 5: Correlations between GMA and Agreeableness and SJT Scores with Theoretically-aligned Combination of Instruction Set and Construct Assessed

	SJT scores when construct & instruction set are aligned ^a	SJT scores when construct & instruction set are <u>not</u> aligned ^b
GMA	↑ .31* ^c	↓ .09 ^c
Agreeableness	↑ .19 ^d	↓ .32* ^d

Note. The vertical arrows represent the posited effects such that ↑ represents an expected stronger relationship and ↓ a weaker one. ^aFor GMA, this would consist of the combination of should-do instruction set with the cognitive construct assessed. For agreeableness, this consists of the combination of the would-do instruction set with the agreeableness construct assessed. ^bFor GMA, this consists of the combination of would-do instruction set with the cognitive construct assessed. For agreeableness, this consists of the combination of the should-do instruction set with the agreeableness construct assessed. ^cA z -test employing Fisher's r -to- Z transformation indicates these two correlations are not significantly different ($p > .05$) ^dA z -test using Fisher's r -to- Z transformation indicates these two correlations are not significantly different ($p > .05$). * $p < .05$, one-tailed.