Comparative Evaluation of Three Situational Judgment Test Response Formats in Terms of Construct-Related Validity, Subgroup Differences, and Susceptibility to Response Distortion

Winfred Arthur Jr., Ryan M. Glaze, Steven M. Jarrett, Craig D. White, and Ira Schurig
Texas A&M University

Jason E. Taylor
People Answers, Inc., Dallas, Texas

As a testing method, the efficacy of situational judgment tests (SJTs) is a function of a number of design features. One such design feature is the response format. However, despite the considerable interest in SJT design features, there is little guidance in the extant literature as to which response format is superior or the conditions under which one might be preferable to others. Using an integrity-based SJT measure administered to 31,194 job applicants, we present a comparative evaluation of 3 response formats (rate, rank, and most/least) in terms of construct-related validity, subgroup differences, and score reliability. The results indicate that the rate-SJT displayed stronger correlations with the hypothesized personality traits; weaker correlations with general mental ability and, consequently, lower levels of subgroup differences; and higher levels of internal consistency reliability. A follow-up study with 492 college students (Study 2; details of which are presented in the online supplemental materials) also indicates that the rate response format displayed higher levels of internal consistency and retest reliability as well as favorable reactions from test takers. However, it displayed the strongest relationships with a measure of response distortion, suggesting that it is more susceptible to this threat. Although there were a few exceptions, the rank and most/least response formats were generally quite similar in terms of several of the study outcomes. The results suggest that in the context of SJTs designed to measure noncognitive constructs, the rate response format appears to be the superior, preferred response format, with its main drawback being that it is susceptible to response distortion, although not any more so than the rank response format.

Keywords: situational judgment tests, response formats, subgroup differences, test taker reactions, noncognitive constructs

Supplemental materials: http://dx.doi.org/10.1037/a0035788.supp

As a predictor method, situational judgment tests (SJTs) are conceptualized as low-fidelity simulations where test takers are presented with work-related situations and a set of predetermined responses (Motowidlo, Dunnette, & Carter, 1990). Although construct explication in the SJT research literature continues to be poor (see Arthur & Villado, 2008; Christian, Edwards, & Bradley, 2010; Schmitt & Chan, 2006), SJTs can be designed to measure a number of constructs (e.g., job knowledge, interpersonal skills, teamwork, leadership, conscientiousness, agreeableness, emotional stability; Christian et al., 2010). Furthermore, like any other method (e.g., assessment centers, interviews), there is a clear recognition that the efficacy of SJTs is influenced by their design features. As such, a number of these features have been investigated, including the modes of presentation and level of fidelity (e.g., written, verbal, video-based, or computer-based; Chan & Schmitt, 2002; Cleverg, Pereira, Wiechmann, Schmitt, & Harvey, 2001; Olson-Buchanan et al., 1998; Weekley & Jones, 1997), response instructions (i.e., behavioral- vs. knowledge-based; McDaniel, Whetzel, Hartman, Nguyen, & Grubb, 2006; Ployhart & Ehrhart, 2002), scoring method (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006), and stem complexity (Ployhart & MacKenzie, 2011). Ployhart and MacKenzie (2011) have presented an informative review and description of these design features.

This article was published Online First February 3, 2014.
Winfred Arthur Jr., Ryan M. Glaze, Steven M. Jarrett, Craig D. White, and Ira Schurig, Department of Psychology, Texas A&M University; Jason E. Taylor, People Answers, Inc., Dallas, Texas.

This research was partially funded by an award from the Texas A&M College of Liberal Arts Cornerstone Faculty Fellowship awarded to Winfred Arthur Jr.

Correspondence concerning this article should be addressed to Winfred Arthur Jr., Department of Psychology, Texas A&M University, 4235 TAMU, College Station, TX 77843-4235. E-mail: w-arthur@tamu.edu
Response Formats

One design feature that has received relatively little research attention is the response format. Three common response formats in the extant literature—the rate, rank, and most/least—are the focus of the present study. The rate format instructs test takers to rate each response option—usually on a 1–5 point scale—in terms of its effectiveness as a response to the situation presented in the scenario (i.e., item stem). The most/least response format instructs test takers to identify the most and least effective options, and the rank response format instructs test takers to rank order the response options from the most effective to least effective.

Although response formats are potentially a critical design feature of SJTs, there is no guidance in the extant literature as to which to use based on their relative superiority on critical psychometric properties and other indicators that are of importance to personnel researchers and practitioners. Indeed, a thorough search of the extant literature resulted in only one study that examined the effects of response formats on the psychometric properties of SJT scores. Specifically, Ployhart and Ehrhart (2002) compared the effects of two response instructions (i.e., “should do” vs. “would do”) and three response formats (i.e., best, best-and-worst, and rate-each-response) on SJT scores. Their results suggest that scores for the best-and-worst response format had lower internal consistency than the rate response format. The criterion-related validity did not appear to vary as a function of response format.

So, given the limited research on the comparative efficacy of different response formats, the objective of the present study is to compare the rate, rank, and most/least response formats. Specifically, in a between-subjects, quasi-experimental field study using a sample of job applicants, the three response formats were compared in terms of construct-related validity, subgroup differences, and score reliability. In addition, to address a potential alternative explanation of the obtained results, a follow-up lab-based experiment, which was a constructive replication and extension of the field study, was implemented. Consequently, this follow-up study (details of which are published in the online supplemental materials) also permitted the assessment of the comparative susceptibility to response distortion, and the internal consistency, test–retest, and alternate-form reliabilities of the scores of the three response formats along with test taker reactions.

To understand how design features can influence test scores and their psychometric properties, it is informative to consider the processes by which test takers respond to test items and response options. Ployhart’s (2006) predictor response process model is most germane to the present study. This model posits that test takers engage in four processes when responding to SJT items—comprehension, retrieval, judgment, and response—all of which are influenced by individual differences and design features. So, for instance, when test takers complete the rate-SJT, they undergo the response process for each response option independently. (It is important to note that the rate permits ties.) However, when test takers complete the rank-SJT, they must make comparative judgments which may require multiple iterations before they arrive at or generate a final response. For example, a test taker may read and comprehend a response option, retrieve pertinent information, and form a tentative judgment. After completing this series of cognitive processes for each response option, the test taker must remember the tentative judgments for each response option and then decide on the relative effectiveness of each option in order to rank them. The complexity and elaborateness of this process (which is similar to that undertaken for the most/least response format) is further magnified by the number of response options. Finally, because the rank-SJT does not permit ties, it (and to a lesser extent the most/least response format) requires test takers to make nuanced distinctions between all response options; hence, having to decide between two or more ostensibly equally effective (or ineffective) options may require a great deal of consideration.

In sum, the rank- and most/least-SJTs require relative judgments across response options and, thus, correspondingly require comparatively higher levels of information processing than the rate-SJT, which is more of an absolute judgment or rating system. As a result, one would also expect different completion times such that ranking the response options should require longer completion times than selecting the most and least effective response, which should in turn require longer times to complete than rating each response. This is commensurate with the proposition that longer response latencies are associated with items that have higher cognitive demands (Bassili, & Scott, 1996; Yan & Tourangeau, 2008).

Construct-Related Validity

Consonant with the fact that SJTs can be designed to measure a wide range of constructs (Arthur & Villado, 2008; Christian et al., 2010; Ployhart & MacKenzie, 2011), the SJT used in the present study was designed to assess an individual’s propensity to engage in counterproductive work behaviors (e.g., theft, damaging company property) that would result in monetary loss to the organization with a particular emphasis on organizations in the retail and hospitality industries. Thus, it was designed to measure the honesty and fidelity aspects of integrity (Berry, Sackett, & Wiemann, 2007), and it accomplished this by asking test takers to indicate how an employee should respond to specified scenarios (see Appendix A for a sample item). It is also important to clarify that the present study is a method- or predictor-design-features study with a focus on the comparative evaluation of three situational judgment response formats in the context of an SJT that is designed to measure a noncognitive construct. Consequently, in terms of its goals and objectives, important boundary conditions for the present study are that (a) the SJT should measure a noncognitive construct and (b) using the same SJT, the three response formats should be manipulated in the comparative evaluation (Arthur & Villado, 2008). With that as a backdrop, we subsequently draw on the integrity literature to develop the conceptual basis for our construct-related validity hypotheses.

Because of the posited differences in information processing and cognitive demands engendered by the different response formats, a point of interest is the relationship between the (integrity) scores obtained for the three SJT response formats and general mental ability (GMA). Ones (1993) reported a sample-weighted mean correlation of .02 between integrity tests and GMA. Consequently, one would not expect a construct-based relationship between the scores on the present SJT and GMA. So, holding the construct constant, any observed variation in the relationships with GMA would be attributable to the differences in the response formats. Under these circumstances—that is, the focal construct is noncognitive and is not theorized to have meaningful relationships
with GMA—the preferred response format would be one that displays the weakest relationship with GMA. A zero relationship is not expected because SJTs by their very nature place unique cognitive demands on test takers (Marentette, Meyers, Hurtz, & Kuang, 2012). However, based on the response process model, it was posited that the rank-SJT requires comparatively higher levels of cognitive and information processing than the most/least-SJT, which in turn requires more cognitive and information processing than the rate-SJT. Consequently, the following was hypothesized:

**Hypothesis 1a:** The rank-SJT will demonstrate stronger correlations with GMA than the most/least-SJT.

**Hypothesis 1b:** The most/least-SJT will display stronger correlations with GMA than the rate-SJT.

In addition, as previously noted, the differential cognitive and information processing demands engendered by the different response formats leads to the proposition that the three response formats will also vary in terms of their completion times. Specifically, because it is well established that longer response latencies are associated with items that have higher cognitive demands (Bassili, & Scott, 1996; Yan & Tourangeau, 2008), it is expected that the rank-SJT will display the longest completion times, and the rate-SJT will display the shortest. So, the following was hypothesized:

**Hypothesis 2:** The rank-SJT will demonstrate the longest completion time, followed by the most/least-SJT, and then the rate-SJT.

The integrity testing literature suggests that agreeableness, conscientiousness, and emotional stability are moderately to strongly related to integrity (Berry et al., 2007; Wanek, Sackett, & Ones, 2003). For instance, Ones (1993) reported mean rs of .26, .28, and .22 for agreeableness, conscientiousness, and emotional stability, respectively—a pattern of results that is consonant with Ones, Viswesvaran, and Schmidt’s (1993) view that integrity is primarily comprised of conscientiousness, agreeableness, and, to a lesser extent, emotional stability. Thus, integrity has an established normative relationship with these personality traits. Hence, one would expect scores on the present SJT—an integrity-based measure—to be related to these personality traits. However, to the extent that the different response formats place different cognitive and information processing demands on test takers, the magnitude of the observed relationships should vary as a function of the response format. Specifically, the cognitive and information processing demands associated with the rank- and most/least-SJTs may be a source of construct-irrelevant variance that, via contamination, may attenuate the relationship between the specified personality traits and the integrity scores measured by the SJT used in the present study. Consequently, the following hypotheses were tested:

**Hypothesis 3a:** The rate-SJT will display stronger correlations with agreeableness, conscientiousness, and emotional stability than the most/least-SJT.

**Hypothesis 3b:** The most/least-SJT will display stronger correlations with agreeableness, conscientiousness, and emotional stability than the rank-SJT.

### Subgroup Differences

An important issue of interest to both researchers and practitioners is the extent to which selection tests display subgroup differences in test scores, which often translate into adverse impact. Consequently, building on the previously discussed construct-related validity propositions, the scores from the three response formats were compared in terms of race- (with a primary focus on White–African American differences) and sex-based subgroup differences. Because integrity tests show negligible White–African American differences (Ones, 1993), it is reasonable to expect negligible race-based subgroup differences on an integrity-based SJT as well. However, it is also recognized that cognitively loaded tests display race-based subgroup differences (Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Specifically, African Americans tend to score lower on cognitively loaded tests than Whites. However, Asian Americans obtain higher scores than Whites (Hough, Oswald, & Ployhart, 2001; McDaniel, Kepes, & Banks, 2011).

Whetzel, McDaniel, and Nguyen (2008) found that SJTs that load higher on cognitive ability produced larger White–African American, and White–Hispanic differences. However, it is noteworthy that for their White–Asian American comparison, Asians also obtained lower scores, a finding that is at odds with the general finding that Asians generally obtain higher scores on GMA tests than Whites (Hough et al., 2001; McDaniel et al., 2011). Similarly, in their review of the literature, Ployhart and Holtz (2008) also found that SJTs (constructs unspecified) show subgroup differences favoring Whites over Asians, African Americans, and Hispanics. In summary, Whetzel et al.’s results indicate that reducing the cognitive load—the extent to which SJT scores are correlated with cognitive ability—may mitigate race-based subgroup differences on SJTs. For example, Chan and Schmidt (1997) found that a video-based SJT (designed to measure work habits and interpersonal skills) resulted in lower White–African American subgroup differences compared to a paper-and-pencil version of the SJT. These results were attributed to the absence of a reading comprehension component in the video-based SJT. Chan and Schmitt’s findings demonstrate the importance of method design features and their influence on the cognitive loading and subsequently, race-based subgroup differences. The response format is an example of another design feature that may influence White–African American subgroup differences via the extent to which the response formats are related to cognitive ability. Hence, it was posited that response formats that engender more cognitive and information processing demands, irrespective of the focal noncognitive construct, will produce comparatively larger White–African American differences than those that do not. Consequently, the following hypotheses were tested:

**Hypothesis 4a:** The rank-SJT will display larger White–African American differences than the most/least-SJT.

**Hypothesis 4b:** The most/least-SJT will display larger White–African American differences than the rate-SJT.

Concerning sex-based subgroup differences, because we had no specific propositions about potential differences, these analyses were primarily exploratory in nature. Nevertheless, females tend to score higher on integrity tests compared to males (ds range from
Research Question 1: Are there sex-based differences on the integrity-based SJT scores?

Research Question 2: Do the rate-, most/least- and rank-SJT display similar levels of sex-based subgroup differences?

Participants

The study sample consisted of 31,194 applicants who completed a GMA test, personality measure, and the integrity-based SJT in an unproctored Internet-based selection battery assessment. It is important to note that although the SJT was administered as part of an operational selection battery, the applicants were informed that it was for test development purposes only. Applicants completed one of three SJTs, with 10,421 completing the rate-SJT (33.41%), 10,435 completing the rank-SJT (33.45%), and 10,338 completing the most/least-SJT (33.14%).

The rollout of the three SJTs was implemented sequentially in the order listed above, with 7 days allotted for each version before it was taken offline. Thus, the data presented here were collected in a 3-week window. The characteristics of the study sample, which are presented in Table 1, indicate an absence of any systematic differences between the three SJT response format groups on the specified study variables—specifically, age, sex, GMA, and the Five-Factor Model (FFM) traits—along with race. The testing firm uses the GMA and personality tests for a wide range of positions for its clients, and although the specific job titles for which each applicant was applying was unavailable to us, applicants applied for a total of over 1,000 different jobs that included position titles such as sales associate, sales specialist, and customer service representative. These positions were in 102 client organizations in the retail and hospitality industries.

Measures

SJT. The SJT was a proprietary untimed 20-item loss prevention measure that assessed an individual’s propensity to engage in counterproductive work behaviors that would result in monetary loss to the organization with a focus on organizations in the retail and hospitality industries. A critical incident approach along with a multiphase iterative subject-matter expert (SME) review process was used to develop, revise, and refine the scenarios and response options. Pilot testing of the items resulted in further refinements and revisions to the items before the data used in the present study were collected. Each item consisted of a description of a situation and five response options.

The test instructions—a third person knowledge-based instruction set—and test items were the same for all test takers. For the rate-SJT, test takers were instructed to rate each response option in terms of its effectiveness on a 5-point scale (5 = very effective). For the most/least-SJT, test takers were instructed to select the most and least effective response option for each item. For the rank-SJT, test takers were instructed to rank the response options.

Table 1

Descriptive Statistics for the Specified Sample Characteristics by the Integrity-Based SJT Response Format

<table>
<thead>
<tr>
<th>Variable</th>
<th>Rate</th>
<th></th>
<th>Rate</th>
<th></th>
<th>Rate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Age</td>
<td>24.87</td>
<td>10.12</td>
<td>24.25</td>
<td>9.52</td>
<td>25.27</td>
<td>10.28</td>
</tr>
<tr>
<td>Integrity-based SJT</td>
<td>60.06*</td>
<td>16.87</td>
<td>53.19</td>
<td>15.44</td>
<td>82.85*</td>
<td>11.96</td>
</tr>
<tr>
<td>GMA</td>
<td>24.80</td>
<td>9.02</td>
<td>24.55</td>
<td>8.93</td>
<td>24.22</td>
<td>8.88</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>77.58</td>
<td>10.12</td>
<td>77.29</td>
<td>10.08</td>
<td>77.56</td>
<td>10.25</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>80.20</td>
<td>9.09</td>
<td>80.09</td>
<td>9.15</td>
<td>80.33</td>
<td>9.24</td>
</tr>
<tr>
<td>Emotional stability</td>
<td>67.68</td>
<td>7.97</td>
<td>67.64</td>
<td>8.11</td>
<td>67.96</td>
<td>8.12</td>
</tr>
<tr>
<td>SJT completion time*</td>
<td>14.35</td>
<td>7.18</td>
<td>20.88</td>
<td>9.09</td>
<td>17.76</td>
<td>8.10</td>
</tr>
<tr>
<td>N</td>
<td>10,421</td>
<td></td>
<td>10,435</td>
<td></td>
<td>10,338</td>
<td></td>
</tr>
<tr>
<td>% female</td>
<td>71.06</td>
<td>.95</td>
<td>70.40</td>
<td>.87</td>
<td>60.97</td>
<td>.91</td>
</tr>
</tbody>
</table>

Note. SJT = situational judgment test; Rate = rate the effectiveness of each option; Rank = rank the options from most to least effective; Most/least = choose the most and least effective options; GMA = general mental ability.

* Rate n = 9,384; rank n = 9,030; and most/least n = 9,345. Rate GMA n = 3,711; rank GMA n = 3,678; and most/least GMA n = 3,747. Standardized mean difference between rate and rank conditions. Standardized mean difference between rate and most/least conditions. Standardized mean difference between rate and most/least conditions; all differences were computed by subtracting the second condition from the first condition. Phi coefficient between rate and rank conditions. Phi coefficient between rank and most/least conditions. Phi coefficient between rate and most/least conditions; SJT and FFM scores range from 0 to 100, and GMA scores range from 0 to 60. Because the absolute magnitude of the means and standard deviations (but not the correlations) are impacted by the specific scoring algorithm used to score the SJTs (see Appendix B), statistical tests for differences between these means are not presented because the results are dependent on the specific scoring algorithm used. p < .05 (two-tailed).
The SJT scoring keys were generated by six SMEs from six different retail-based client companies of the assessment firm. These SMEs were either human resource executives, directors, managers, or loss prevention experts or specialists in their respective organizations. For the rate-SJT, the SMEs independently rated each response option in terms of its effectiveness on a 5-point scale (5 = very effective). The SME ratings were aggregated and the keyed responses were the average of the SME ratings (rounded to the whole number using standard rounding rules). For the rank-SJT, the SMEs independently ranked (without ties) the response options from the most to the least effective. The average SME ranking within each item was used to designate the keyed rank order. Finally, for the most/least-SJT, the highest and lowest ranked response options from the rank-SJT scoring key were designated as the most and least effective response options, respectively.

The agreement amongst the SMEs in generating the scoring keys was generally quite high. Specifically, the mean rank-SJT scoring key was .84 (Mdn = .92, SD = .26, minimum = .00, maximum = 1.00) for all response options. Likewise, the mean for the ranking scoring key was .87 (Mdn = .92, SD = .21, minimum = .00, maximum = 1.00). For the most/least scoring key, the mean rwg for the response options identified as the least effective was .89 (Mdn = 1.00, SD = .24, minimum = .00, maximum = 1.00); and for the most effective, it was .96 (Mdn = 1.00, SD = .09, minimum = .60, maximum = 1.00). Given some variability amongst the SMEs for some of the items, this could have been addressed and mitigated via a consensus meeting instead of just mathematically aggregating the SMEs’ ratings and rankings. However, in the present situation, a consensus meeting was not available as an option. So the empirical effects of the low agreement for the specified items were assessed by dropping the four items for which there was no agreement amongst the SMEs (i.e., rwg = .00). The correlations between the resultant 16-item and 20-item tests were 1.00, .99, and .99 for the rate, rank, and most/least response formats, respectively. Furthermore, the magnitude and pattern of the results for the 16-item and 20-item tests were identical.

In scoring the SJTs, for the rate response format, test takers were awarded a point for each response rating that matched the SME rating. Consequently, test takers could receive from 0 to 5 points for each item, and test scores could range from 0 to 100. For the rank, test takers received a point for each response ranking that matched the SME ranking. Thus, test takers could receive from 0 to 5 points for each item, and test scores could range from 0 to 100. For the most/least, as per Motowidlo et al. (1990; see also Ployhart & MacKenzie, 2011), test takers were awarded a point each for correctly identifying the most and the least effective response options. Test takers were docked a point each for identifying the most effective response option as the least effective, and the least effective response option as the most effective. Test takers received no points for identifying a non-keyed response option as the most or least effective. Consequently, for each item, test takers’ scores could range from −2 to 2. Because the test score was the sum of item scores across the 20 scenarios, the resultant SJT test scores could range from −40 to 40. However, for ease of interpretation and comparative purposes, the test scores were first scaled to be positive (i.e., 0–80), and then converted to 0–100 points to match the score range of the other keys.

Different scoring methods can obviously influence the magnitude and range of obtained test scores. However, the critical issue in terms of the present comparative evaluation is whether they also alter the rank order of test takers. Consequently, the effect of alternative scoring methods on the obtained results was investigated. The results of these alternative scoring methods, which are presented in Appendix B, indicate that there were negligible differences in both the pattern and magnitude of the results.

**GMA test.** The GMA test was a proprietary unproctored Internet-based speeded test that consisted of 60 four-alternative multiple-choice items with verbal (28 items) and numeric (32 items) subscales. Participants had 10 min to complete the test. Arthur, Glaze, Villado, and Taylor (2010) reported a retest reliability coefficient of .78 (mean retest interval = 429.16 days, SD = 54.84) for an equivalent form of this test, along with a convergent validity of .72 with the Thurstone Test of Mental Alertness (Thurstone & Thurstone, 1952). The assessment firm administered the GMA test to only a subset of applicants, depending on the specific organization and job to which the test taker was applying (i.e., for some job applicants, the GMA test was not part of the selection test battery). Thus, out of the study sample of 31,194 applicants, 11,136 (35.70%) completed the GMA test, with 3,711 (35.61%) for the rate-SJT, 3,678 (35.25%) for the rank-SJT, and 3,747 (36.24%) for the most/least-SJT. Consequently, GMA-related analyses are limited to this smaller subset of the sample, while all other analyses used the total sample. It should be noted that identical results (in terms of pattern, magnitude, and statistical significance) were obtained when a casewise deletion approach (that limited all analyses to only cases for which there were complete data) was used. The results of these analyses are available from the first author.

**Personality measure.** The personality test was a proprietary unproctored Internet-based untimed measure that consisted of 22 10-item personality scales. These scales were previously sorted into the FFM dimensions by Arthur et al. (2010), who used procedures, guidelines, and information presented in Barrick and Mount (1991) as well as Birkeland, Manson, Kisamore, Brannick, and Smith (2006). Internal consistency reliability estimates for the scores of the three dimensions used here were .88, .96, and .75 for agreeableness, conscientiousness, and emotional stability, respectively.

**Results**

Comparisons of characteristics of the samples for the three response format conditions are presented in Table 1 along with the descriptive statistics for the response formats. The scores for all the SJTs displayed high levels of internal consistency, with the rate-SJT having the highest reliability estimate (.95), followed by the most/least-SJT (.91), and then the rank-SJT (.87). Table 2 presents the correlations between GMA, agreeableness, conscientiousness, and emotional stability, and the SJT scores for each response format.
format. The correlations between SJT scores and GMA were .28, .25, and .16, for the rank-, most/least-, and rate-SJTs, respectively. The correlation between the rank-SJT scores and GMA was stronger than that for the most/least-SJT scores (z = 2.16, p < .05), thus, supporting Hypotheses 1a. Furthermore, the correlation between GMA and the most/least-SJT was stronger than that of the rate-SJT (z = 6.48, p < .05), supporting Hypothesis 1b.

Tests of Hypothesis 2 provided additional support for the differential cognitive and information processing demands proposition that the three response formats displayed different completion times. As hypothesized, the rank displayed the longest mean completion time (M = 20.83, SD = 9.09; n = 9,030), which was significantly longer than that for the most/least-SJT (M = 17.76, SD = 8.10; n = 9,345; d = 0.36, p < .05), which was in turn longer than that for the rate-SJT (M = 14.35, SD = 7.18; n = 9,384; d = 0.45, p < .05). In summary, these differential completion times (reported above in minutes) are consonant with the differential cognitive and information processing demands engendered by the three response formats.

As posited by Hypothesis 3a, agreeableness, conscientiousness, and emotional stability demonstrated stronger relationships with rate-SJT scores compared to the most/least-SJT scores (z = 7.09, 8.78, and 7.40 for agreeableness, conscientiousness, and emotional stability, respectively; all differences were statistically significant, p < .05). The most/least-SJT showed weaker relationships with the specified personality traits compared to the rank-SJT, but the differences were not statistically significant (z = 1.50, 0.75, and 0.73, p > .05, for agreeableness, conscientiousness, and emotional stability, respectively). Thus, Hypothesis 3b, which stated that the most/least-SJT scores would display stronger relationships with the specified personality traits compared to the rank-SJT scores, was not supported.

Table 3 presents the race-, and sex-based, subgroup differences for each response format. The results show that the rank and most/least response formats displayed similar subgroup differences when African American and Asian American applicants were compared to Whites. Thus, the data did not support Hypothesis 4a, which stated that the rank-SJT would display larger race-based subgroup differences compared to the most/least-SJT. However, both the rank- and the most/least-SJT displayed larger race-based subgroup differences than the rate-SJT, thus providing support for Hypothesis 4b. Interestingly, it is worth noting that the lower Asian American scores, compared to Whites, obtained here is the same pattern of results that was obtained by Whetzel et al. (2008) in their SJT–cognitive load meta-analysis and by Ployhart and Holtz (2008) in their review of SJTs.

Concerning sex-based differences, women displayed higher scores on all three response formats. At the construct-level, these findings are consistent with the extant literature (e.g., Berry et al., 2007) which indicates that women generally score higher on integrity tests than men. The results also indicate that the rank-SJT resulted in a sex-based subgroup difference (d = −0.21) that was larger than that for the most/least-SJT (d = −0.09, z = 194.89, p < .05) and the rate-SJT (d = −0.08, z = 197.88, p < .05). The difference between the rate-SJT and most/least-SJT was also significant (z = 16.14, p < .05).

### Discussion

Given the relatively large sample sizes, it is recognized that even fairly small effects are likely to be statistically significant. Consequently, the focus of the following discussion is more on the magnitude and pattern of results and less so on their statistical significance. That being said, a number of summary statements can be made on the basis of the results. First, the highest internal consistency reliability was obtained for the rate-SJT scores (.95), followed by the most/least-SJT (.91), and then the rank-SJT (.87). Second, the posited higher cognitive and information processing demands associated with the rank- and most/least-response formats is reflected in their higher relationship with GMA, and the differential completion times as well. Consequently, our findings also have implications for incremental validity where one wants measures that are each related to the criterion but not to each other. Specifically, the use of a rate response format for a noncognitive SJT is likely to result in higher incremental validity over GMA than a rank or most/least response format for the same SJT.

Because integrity tests generally display meaningful relationships with agreeableness, conscientiousness, and emotional stabil-

---

**Table 2**

**Integrity-Based SJT Correlations With General Mental Ability and the Specified Five-Factor Model Personality Traits for All Response Formats**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Rate</th>
<th>Rank</th>
<th>Most/least</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMA</td>
<td>.16</td>
<td>.28</td>
<td>.25</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.31</td>
<td>.20</td>
<td>.18</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.34</td>
<td>.19</td>
<td>.18</td>
</tr>
<tr>
<td>Emotional stability</td>
<td>.24</td>
<td>.11</td>
<td>.10</td>
</tr>
</tbody>
</table>

*Note. All correlations are statistically significant at p < .01 (one-tailed). SJT = situational judgment test; Rate = rate the effectiveness of each option (n = 10,421); Rank = rank the options from most to least effective (n = 10,338); Most/least = choose the most and least effective options (n = 10,338); GMA = general mental ability. Ones (1993) reported meta-analytic sample-weighted mean correlations between integrity (tests) and GMA (.02), agreeableness (.26), conscientiousness (.28), and emotional stability (.22).*
and least effective options. SJT scores range from 0 to 100. Differences (Ones, 1993), unlike the rank- and most/least-SJTs, the tests tend to display small to negligible race-based subgroup differences. Consonant with the extant finding that traditional integrity GMA (.28 and .25, respectively) than the rate (.16).

The increased cognitive and information processing demands—a pattern of results that were nomologically consonant with the extant literature (see Table 2). In contrast, weaker relationships were obtained for the rank and most/least response formats—a pattern of results which is consistent with the increased cognitive and information processing demands associated with these two formats. This is further reflected in the finding that the rank and most/least response formats had higher relationships with GMA (.28 and .25, respectively) than the rate (.16).

The increased cognitive and information processing demands may also account for the observed race-based subgroup differences. Consonant with the extant finding that traditional integrity tests tend to display small to negligible race-based subgroup differences (Ones, 1993), unlike the rank- and most/least-SJTs, the rate-SJT displayed race-based subgroup differences commensurate with that of traditional integrity tests. It is also worth noting that while the White–African American findings are at odds with the observation that the White–Asian GMA score difference tends to favor Asians (Hough et al., 2001; McDaniel et al., 2011). Additional SJT construct-level studies are warranted to shed further light on this.

In summary, the pattern and magnitude of the results are supportive of the summary conclusion that in the context of noncognitive constructs, because of its lower cognitive and information processing demands, the rate response format may be preferable to the rank and most/least response formats particularly when the goal is to minimize the g-loading of noncognitive SJTs, especially in instances where the assessment system already includes an ability or g-loaded test. This conclusion notwithstanding, because the rate-SJT shared a similar response format as the personality measure (i.e., a Likert-type response format), and the rank- and most/least-SJTs shared what appears to be a similar response format with the GMA test (i.e., a multiple-choice-type response format), a shared-common-response-method effect (Podsakoff, MacKenzie, & Podsakoff, 2012) may serve as a plausible alternative explanation for the observed effects instead of the posited differences in the cognitive and information-processing demands.4

From an applied practical perspective, the import of a shared-common-response-method effect may not be as pivotal since from this perspective, it is difficult to envisage either a personality measure that has a response format similar to a GMA (or knowledge) test (i.e., multiple-choice) or a GMA (or knowledge) test with a response format similar to a personality measure (i.e., Likert-type response scale). Furthermore, in support of our conceptual explanation for the observed effects, both the completion time and the subgroup differences obtained in the present study also align with the supposition that the response formats engendered varying levels of cognitive and information-processing demands.

Note. All $d$s are statistically significant at $p < .01$ (one-tailed). For race, the other subgroups are compared to White participants, such that a positive $d$ indicates that White participants scored higher than the subgroup to which they are compared. Likewise for sex, a positive $d$ indicates that males scored higher than females. For age, a positive $d$ indicates that test takers less than 40 years of age scored higher than test takers 40 years of age or older. SJT = situational judgment test; Rate = rate the effectiveness of each option; Rank = rank the options from most to least effective; Most/least = choose the most and least effective options. SJT scores range from 0 to 100.

<table>
<thead>
<tr>
<th>Response format</th>
<th>Race</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>African American</td>
<td>Asian American</td>
<td>Hispanic</td>
<td>White</td>
<td>Male</td>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>1,918</td>
<td>571</td>
<td>1,516</td>
<td>5,537</td>
<td>2,998</td>
<td>7,405</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>57.12</td>
<td>56.03</td>
<td>58.76</td>
<td>61.84</td>
<td>58.69</td>
<td>60.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SD$</td>
<td>19.57</td>
<td>19.48</td>
<td>17.60</td>
<td>15.01</td>
<td>18.04</td>
<td>16.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d$</td>
<td>0.29</td>
<td>0.38</td>
<td>0.20</td>
<td></td>
<td>18.04</td>
<td>16.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>1,925</td>
<td>533</td>
<td>1,601</td>
<td>5,503</td>
<td>3,072</td>
<td>7,346</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>47.73</td>
<td>50.16</td>
<td>49.72</td>
<td>56.34</td>
<td>50.87</td>
<td>54.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SD$</td>
<td>16.52</td>
<td>15.11</td>
<td>15.97</td>
<td>14.16</td>
<td>15.88</td>
<td>15.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d$</td>
<td>0.58</td>
<td>0.43</td>
<td>0.47</td>
<td></td>
<td>15.88</td>
<td>15.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Most/least</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>1,845</td>
<td>572</td>
<td>1,664</td>
<td>5,389</td>
<td>3,043</td>
<td>7,283</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$M$</td>
<td>62.83</td>
<td>63.72</td>
<td>64.67</td>
<td>68.13</td>
<td>65.67</td>
<td>66.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SD$</td>
<td>12.89</td>
<td>11.49</td>
<td>10.59</td>
<td>7.11</td>
<td>9.49</td>
<td>9.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d$</td>
<td>0.59</td>
<td>0.58</td>
<td>0.43</td>
<td>0.47</td>
<td>0.43</td>
<td>0.47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4 This issue was raised by a reviewer, and the follow-up lab experiment was designed and implemented specifically to address this concern.
differences is the g-loading of the test or measure, the differences-in-g-loading explanation would predict the pattern of subgroup differences obtained here, whereas a common-response-method effect explanation would not predict or account for the observed subgroup differences. As such, the subgroup differences are more commensurate with the former explanation than they are with the latter.

Nevertheless, in the absence of a design that controls for the response formats by holding them constant across all the measures, the reasoning that the rate-SJT–FFM personality traits relationships and the rank-SJT–GMA relationship are an artifact of the shared common response method and not due to the posited differences in g-loading cannot be fully discounted as an alternative explanation. So, a lab-based experiment using a sample of 492 college students was undertaken to address this concern. This experiment crossed the three response formats with the measurement of all the focal constructs, that is, the integrity-based SJT, and measures of GMA and the specified FFM personality traits. Consequently, if the shared-common-response-method explanation best accounts for the observed response format effects, then the highest positive relationships should be obtained for the matched response formats (e.g., rank-SJT/rank-GMA, rate-SJT/rate-GMA, and most/least-SJT/most/least-GMA) compared to the other (mismatched) conditions. On the other hand, if the differences-in-g-loading explanation best accounts for the results of the study, then said results should be replicated such that the rank-SJT should display the strongest positive relationships with GMA regardless of the GMA response format, and the rate-SJT should display the strongest positive correlations with the specified FFM personality traits regardless of the personality measure response format. The results of the follow-up lab experiment—details of which are reported in the online supplemental materials (as Study 2)—failed to provide support for the shared-common-response-method explanation.

The follow-up study also provided an opportunity to address additional questions pertaining to response distortion, score reliability, and test taker reactions. So, as detailed in the online supplemental materials, the results of the follow-up experiment indicated that the rate-SJT was the most susceptible to response distortion and the most/least-SJT the least; the rank-SJT results were similar to those for the rate-SJT. The susceptibility of the rate-SJT to response distortion leads one to suspect that compared to the other response formats, it may also be more susceptible to coaching as well (Cullen, Sackett, & Lievens, 2006). Concerning test taker reactions, the rank-SJT engendered the least favorable reactions; the rate response format displayed comparatively more favorable test taker reactions compared to the rank. Finally, the rate-SJT scores demonstrated the highest levels of test–retest reliability, and again, the rank-SJT the least. However, the rank- and most/least-SJTs demonstrated the highest alternate-form reliability, and the rate-SJT displayed the lowest correlation with the other two response formats—a pattern of results that is consonant with the proposition that the rank and most/least response formats engender similar levels of cognitive load and, consequently, would be expected to display the highest inter-form correlation.

Limitations and Directions for Future Research

The absence of criterion data did not permit an assessment of the comparative criterion-related validity of the three SJT response formats. However, given the differences in psychometric properties that were obtained in the present study, it would not be unreasonable to expect differences in criterion-related validities as well (e.g., see Lievens & Sackett, 2006). However, this is ultimately an empirical issue, and an investigation of the differential effects of response formats on criterion-related and incremental validity is clearly an avenue of future research.

The results reported here for the rank-SJT must also be interpreted in the context of the fact that we used a five-option SJT. Hence, one can envisage a situation where four- or maybe even a three-option SJT (e.g., see Edwards, Arthur, & Bruce, 2012) might be considered to be less difficult and engender less information processing and cognitive demands than a five-response format. In addition, the rank-SJT may also be influenced by the structural dispersion or dissimilarity of the response options (i.e., the extent to which the relative effectiveness of the options are similar or dissimilar). These are additional design features that are worthy of further research.

In addition, future research that investigates the joint effects of multiple SJT design features would be informative. Such research may assess the comparative efficacy of the response formats by crossing them with different levels of other design features such as modes of presentation (e.g., written vs. verbal vs. video-based), response instructions (e.g., behavioral-based vs. knowledge-based), and the number of response options, and might investigate additional outcomes such as susceptibility to coaching as well.

Implications and Conclusion

The results of the present study indicate that the most/least and rank response formats appeared to introduce g-related construct-irrelevant variance into the integrity-based SJT scores. Ones (1993) reported a GMA-integrity sample-weighted mean r of .02—a relationship that is negligible at best. So, whereas GMA may be job-relevant (i.e., it is relevant to most, if not all, jobs), it is construct-irrelevant in the context of measuring integrity. Consequently, minimizing the construct overlap between a noncognitive SJT by using the rate response format will likely increase the incremental validity of the noncognitive SJT’s construct over GMA when they are both used in a multi-predictor selection battery. Consequently, our results highlight the importance of paying particular attention to the constructs measured by SJTs and, in addition, being cognizant of how SJT (and other predictor method) design features can influence the extent to which they are measuring the specified construct of interest. Aligning the focal construct with response formats and other SJT design characteristics (e.g., response instructions, mode of presentation) is commensurate with a construct focused approach in the comparative evaluation and use of method-based predictors (Arthur, Day, McNelly, & Edens, 2003; Arthur & Villado, 2008; Christian et al., 2010; Huffcutt, Conway, Roth, & Stone, 2001).

In conclusion, organizations should be cognizant of the effects of the three SJT response formats and choose that which is commensurate with their test use and goals. The objective of the present study was to present a comparative evaluation of three SJT response formats in terms of score reliability, construct-related
validity, and levels of subgroup differences—as well as susceptibility to response distortion, score reliability, and test taker reactions in Study 2. Because of its lower information processing demands, the rate-SJT displayed weaker relationships with GMA, stronger effects with the specified personality traits, and smaller subgroup differences compared to the rank and most/least response format. Furthermore, the rate-SJT scores also displayed the highest internal consistency and test–retest reliability estimates, and its level of susceptibility to response distortion was comparable to that for the rank-SJT. It also engendered the most favorable test taker reactions. So, in closing, the pattern and magnitude of the results of the present study are supportive of the summary conclusion that because of its lower cognitive and information processing demands, in the context of noncognitive constructs, the SJT rate response format appears to be preferable to the rank and most/least (at least in terms of the outcomes investigated here) when the goal is to minimize the g-loading of noncognitive SJTs particularly in instances where the assessment battery already includes an ability or g-loaded test.

References


and developing members for the organization (pp. 237–252). doi: 10.1037/12170-008

**Appendix A**

**Sample Situational Judgment Test Item With Rate Response Format**

An employee is hired into a situation where it is common for all employees to take things from work. Rate the effectiveness of each of the following in terms of what you think a new hire should do.

1. Take something small to be sure they fit into the work culture.
2. Take products from work only when they really need extra cash.
3. Write an anonymous letter notifying the corporate office of the situation.
4. Take things and give them to friends and family members to help them out.
5. Work through other employees to sell products and split the cash.

**Rank Response Format Instructions for the Same Item**

An employee is hired into a situation where it is common for all employees to take things from work. Rank the effectiveness of each of the following in terms of what you think a new hire should do.

**Most/Least Response Format Instructions for the Same Item**

An employee is hired into a situation where it is common for all employees to take things from work. Select the most and least effective from the following in terms of what you think a new hire should do.

(Appendices continue)
Appendix B
Results of Alternative Scoring Algorithms

The scoring algorithms used in the present study represent a subset of possible scoring approaches. To investigate the extent to which different scoring methods influenced the observed situational judgment test (SJT) scores, alternative scoring methods were applied to test takers’ responses. Specifically, the most/least-SJT used a partial credit scoring method that is common in the literature (Motowidlo, Dunnette, & Carter, 1990). However, the scoring algorithms used for the rate- and rank-SJTs used an absolute scoring method with no partial credit. Thus, partial credit scoring methods were developed for the rate- and rank-SJTs, and an absolute scoring method was developed for most/least-SJT.

For the rate-SJT partial credit scoring method, the absolute difference between the subject-matter experts’ (SMEs’) ratings and the test takers’ responses was calculated for each response option. The absolute deviation was summed across all response options for all items. The partial credit scoring method for the rank-SJT was computed in the same manner. For the most/least-SJT, test takers were awarded a point for correctly identifying the most and least effective response option (as identified by the SMEs). However, test takers were not docked a point for identifying the most effective response option as the least effective response option, or for identifying the least effective response option as the most effective option.

The results, which are reported in Table B1, indicate that differences in partial versus absolute scoring methods were trivial, at least in the context of this specific SJT. The correlation between the two scoring methods for each specified response format was very large. Indeed, the magnitude of these correlations is in the order of very strong reliability estimates. Interestingly, the most/least-SJT displayed similar means and standard deviations regardless of the scoring method used.

These alternate scoring algorithms were applied to the Study 2 data as well, and a similar pattern of results was obtained.

Table B1
Integrity-Based SJT Score Descriptive Statistics Using Different Scoring Algorithms and Their Intercorrelations

<table>
<thead>
<tr>
<th>Response format</th>
<th>Partial credit</th>
<th>Absolute</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Rate</td>
<td>58.75</td>
<td>44.88</td>
</tr>
<tr>
<td>Rank</td>
<td>72.68</td>
<td>40.45</td>
</tr>
<tr>
<td>Most/least</td>
<td>82.85</td>
<td>11.96</td>
</tr>
</tbody>
</table>

Note. SJT = situational judgment test; r = correlation between scores based on the partial credit scoring algorithm and the absolute scoring algorithm; Rate = rate the effectiveness of each option; Rank = rank the options from most to least effective; Most/least = choose the most and least effective options.

Received August 3, 2011
Revision received November 8, 2013
Accepted December 11, 2013

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at http://notify.apa.org/ and you will be notified by e-mail when issues of interest to you become available!