Optimal Number of Questionnaire Response Categories: More May Not Be Better

W. Paul Jones¹ and Scott A. Loe¹

Abstract
Two multi-instrument investigations in a university clinic/lab provided an opportunity to explore the impact of reducing the number of response alternatives in a scale measuring vocational personality traits. In a simulation study, a standard computer-based administration provided a numeric scale for each item ranging from 0 to 10. The tests were then rescored to simulate the effect of only three choices. For the follow-up study, two versions of the scale were created, one with two response options and the other with six response options, and were randomly assigned to participants. Typical relationships were evident between the vocational personality traits and scores on standard measures of core personality traits with negligible impact from reducing the number of response alternatives. Neither the ability of the participants nor the self-reported distress had a measurable impact on the utility of the results with the reduction in response alternatives.

Keywords
behavioral sciences, educational research, education, social sciences, careers, educational measurement & assessment, educational psychology & counseling, measurement and scaling methods, research methods

The optimal number of response options to maximize reliability and validity in affective assessment scales is a long-standing debate. Bloom, Fischer, and Orme (2003), for example, noted that available research suggests five to seven response categories but add that a 9-point scale may be advantageous if the user is capable of making the distinctions. Krieg (1999) described statistical adjustments to compensate for problems with coarse measurement scales, concluding that increasing the number of response options would be a better choice.

In a general summary of the literature in questionnaire design, Lietz (2010) reported that 5 to 7 response options were most commonly used and that enhanced psychometric qualities, including reliability and validity, were suggested as the justification for including multiple response categories. Rodgers, Andrews, and Herzog (1992) investigated the effect of using 2 to 10 response categories and concluded that expected values of validity coefficients increased by approximately .04 with each additional response option. In contrast, Bendig (1954) and Mattell and Jacoby (1971) studied the effect of using response categories ranging from 2 to 9 and found negligible impact on reliability when the number of response categories was increased.

While Cook, Heath, Thompson, and Thompson (2001) noted that most studies appear to support the desirability of a larger number of response categories, they also cautioned that when more score intervals are available, there is an implicit assumption that the participant has the cognitive capability to process the meaning of the intervals. When that assumption is untenable, the outcome, in addition to a longer administration time, can be a reduction in score reliability. In the background of discussions about the optimal number of response categories there are two other questions. One is the long-standing debate about whether these scales produce only ordinal data that do not support parametric statistical analysis. Jamieson (2004), for example, succinctly argued for ordinal interpretation, noting that the average of “fair” and “good” is not “fair-and-a-half.” It is, however, a common practice for researchers to assume that Likert categories are “close enough” to interval-level measurement, consistent with Blaikie’s (2003) position that interval-level statistics can be justified if used with caution. Knapp (1990) summarized the pros and cons of ordinal and interval assumptions in scales using Likert categories, including challenging a belief that more response categories automatically enhance the appropriateness of parametric analysis.

¹University of Nevada, Las Vegas, USA

Corresponding Author:
W. Paul Jones, Department of Educational Psychology and Higher Education, University of Nevada, 4505 S. Maryland Parkway, Las Vegas, NV 89154-3003, USA.
Email: jones@unlv.nevada.edu
In addition to questions about whether Likert response categories represent equal intervals, there is an underlying issue of inconsistency in category interpretation between individuals contributing to misleading results when responses from research participants are aggregated for analysis. To illustrate, on a 4-point scale ranging from 1 = highly unsatisfied to 4 = highly satisfied, Participant A marks 3 (satisfied) and Participant B marks 4 (highly satisfied). Arguably, the two participants could in fact be expressing exactly the same degree of experience or product satisfaction with the difference in their responses reflecting an idiosyncratic interpretation of the response category. It would be reasonable to assume that what it takes for Participant B to feel “highly satisfied” would be consistent on other questions or on the same question over time. But combining the responses of Participant A and Participant B as a part of a group measure of central tendency is inherently flawed regardless of whether the central tendency is reported as a median (ordinal) or as a mean (interval).

Little attention has been given to the potential for flawed aggregation in the research literature. It would, in fact, be quite difficult to design a study to address this concern without some direct, perhaps physiological, indicator of underlying satisfaction. While it is generally assumed that increasing the number of response categories in Likert items enhances the psychometric properties of a scale, it is possible that doing so simply increases the likelihood that psychologically homogeneous response categories are arbitrarily divided. Stated in the opposite direction, reducing the number of response categories could have a positive effect on validity of aggregated data by reducing the number of opportunities for inappropriately treating different perceptions of response categories as if they were the same.

Data obtained in two multi-instrument investigations in a university clinic/lab provided the opportunity for further investigation of the effect on psychometric characteristics of a scale when the number of alternative responses is varied. First, a simulation study was conducted comparing the effect of using 3 and 10 response categories on a visual analog scale of vocational personality traits. This was followed by a study in which participants were randomly assigned to complete one of two versions of the same vocational personality trait measure, one version with only 2 response categories or the other version with 6 response categories. The studies were designed to address two primary research questions:

*Research Question 1*: Is there a significant impact on measurement statistics contingent on the number of response options?

*Research Question 2*: Is there a significant difference in the prediction of related external criteria contingent on the number of response categories?

These questions were selected, in part, with the belief that unless there were marked differences in favor of the larger number of response categories, researchers might be better served with questionnaires using fewer categories and thus reducing the influence of inter-individual inconsistencies in category interpretation on the resulting statistical estimates.

### Simulation Study

**Method**

**Participants.** Data for the simulation study were provided by a total of 77 undergraduate students attending an urban, southwestern university in the United States. Participants were volunteers who chose this study from several different projects available to meet a research requirement for undergraduate courses in educational psychology. The majority (77%) were female. Most (73%) were in the age range of 18 to 24 with 19% in the age range of 25 to 34, 5% in the age range of 35 to 44, and 3% in the 45 and above age range. Ethnic backgrounds were as follows: Caucasian (66%), African American (10%), Hispanic American (6%), Asian American (5%), and Other or no report (13%). The primary language of the participants in the simulation study was English (92%). The range of self-reported grade point average of the participants on a 4.0 scale was 3.5 to 4.0 (39%), 3.0 to 3.4 (42%), 2.5 to 2.9 (16%), and 2.0 to 2.4 (2%).

**Instrumentation.** *CogStyle*, based on an artificial intelligence model postulated by Lowen (1982), is a scale designed to facilitate equitable assessment of personality traits among persons with and without a disabling visual condition (Jones, 1996). The items provide direct assessment of the six vocational personality traits identified by Holland (1973, 1997): Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. The item format uses paired comparisons of six adjectives, one for each for the six Holland dimensions, and then paired comparisons of six action verbs, each also corresponding to one of the Holland dimensions. Adjectives are practical, curious, flexible, sympathetic, ambitious, and efficient. The actions are fixing things, investigating things, designing things, assisting others, persuading others, and organizing things.

In addition to raw scores on each dimension, results on the Holland scales are typically reported as two- or three-letter codes, indicating the strongest preferences. For example, an IE code indicates that the person’s highest score was on the Investigative scale and the next highest score was on the Enterprising scale, and this ranking then becomes the basis for career exploration. The use of the ranking procedure enables some expected comparability among different vocational interest scales that differ in length and/or type of stimuli used for the assessment and does not require normative transformation.

The NEO Five Factor Inventory (NEO-FFI; Costa & McCrae, 1992) is a 60-item self-report questionnaire measuring each of the “big-five” trait dimensions of
Table 1. Simulation Study: Descriptive Statistics for Vocational Personality Trait Scores With 10-Option and Simulated 3-Option Scenarios (N = 77).

|                      | Realistic | Investigative | Artistic | Social   | Enterprising | Conventional |
|----------------------|-----------|---------------|----------|----------|--------------|--------------|
|                      | M (SD)    | M (SD)        | M (SD)   | M (SD)   | M (SD)       | M (SD)       |
| 10-Option            | 4.4 (1.23) | 4.8 (1.24)    | 5.3 (1.09)| 5.5 (1.39)| 4.4 (1.47)   | 5.5 (1.53)   |
| 3-Option: 1          | 4.3 (1.82) | 4.8 (1.75)    | 5.5 (1.66)| 5.7 (1.90)| 4.0 (2.15)   | 5.7 (1.09)   |
| 3-Option: 2          | 4.4 (1.71) | 4.7 (1.62)    | 5.4 (1.62)| 5.7 (1.75)| 4.2 (2.04)   | 5.7 (1.95)   |

Note: Statistically significant difference between mean scores only on Enterprising scale. 3-Option: 1 is Scenario 1 with original scores of 5 coded as 2 in 3-point simulation. 3-Option: 2 is Scenario 2 with original scores from 4 to 6 coded as 2 in 3-point simulation.

Neuroticism, Extraversion, Openness, Conscientiousness, and Agreeableness. Each scale consists of 12 items about typical behaviors or reactions that are rated according to a 5-point Likert scale, ranging from “strongly disagree” to “strongly agree.” The questionnaire yields T-scores for each of the five trait dimensions. Reliabilities reported in the examiner’s manual for the five dimensions range from .86 to .95.

The Woodcock–Johnson Tests of Cognitive Ability (WJ-III; Woodcock, McGrew, & Mather, 2001) is an individually administered measure of general intellectual ability and specific cognitive abilities in children, adolescents, and adults, aged 2 years through 94 years. The battery yields a General Intellectual Ability index (GIA) that is a measure of g along with cluster scores representing the broad Cattell-Horn-Carroll (CHC) abilities and individual subtest scores.

Procedure. A series of studies conducted in a university clinic/lab in 2005-2006 that included a component of computer-based test administration provided an adventitious opportunity to investigate the effect of the number of response options in a visual analog scale. The 30 items assessing the Holland dimension in the CogStyle scale were adapted to conform with software requirements, resulting in a response mode in which the participant expressed preference in the paired comparisons of adjectives and action verbs on a 0 to 10 visual analog scale. The total score for each dimension was weighted by the extent of preference.

To illustrate, one of the adjectives for the Realistic score is “practical.” One of the adjectives for the Investigative score is “curious.” The item was presented on the computer screen to participants with “practical” on the 0 end of the visual analog scale and “curious” on the 10 end of the scale with instructions for participants to indicate their preference. If the participant clicked the 0, one point was added to the Realistic score; if the participant clicked the 10, one point was added to the Investigative score. If the participant clicked the 1, .9 was added to the Realistic score; if the participant clicked the 9, .9 was added to the Investigative score, and so forth. Participants who clicked the mid-point, 5, had .5 added to the Realistic and to the Investigative score.

To estimate the effect of a significant reduction in response options, the scales were then rescored with two scenarios to simulate responses if there had been only three options to express the degree of preference. For the first simulation scenario, using the “practical” or “curious” example in the previous paragraph, clicking 0 to 4 resulted in one point added to the Realistic score; clicking 6 to 10 resulted in one point added to the Investigative score. Clicking the mid-point of the scale resulted in .5 point added to each of the two scores.

Recognizing that it would be at least equally plausible to extend the range for the simulated middle score, a second 3-option scenario used a different rubric. In the second scenario, when simulating the scoring of the “practical” or “curious” stimulus, responses from 0 to 3 added one point to the Realistic score, responses from 7 to 10 added one point to the Investigative score, responses of from 4 to 6 added .5 to each of the two scores, and so forth.

An appointment session for each participant began with the completion of demographic information and review and signature on human subjects consent forms. Individual administration of selected cognitive scales was then provided by school psychology graduate students, followed by completion of a series of computer-administered scales, including the vocational personality scale that is the focus of this study.

Results of the Simulation Study

Table 1 displays the descriptive information for the 10-option and the two simulated 3-option scenarios for each of the six vocational personality scales. A perusal of Table 1 suggests that the different response options did not appear to have a marked effect on the obtained mean scores. Paired sample t test analyses found three statistically significant differences in the mean scores obtained from the scoring alternatives with Bonferroni correction for the 18 comparisons. Two of the statistically significant differences were on the Enterprising scale: 10-option versus 3-option-scenario 1, t(76) = 3.949, p = .000, and 10-option versus 3-option-scenario 2, t(76) = 3.402, p = .001. There was one statistically significant difference on the Conventional scale: 10-option versus 3-option-scenario 2, t(76) = 3.353, p = .001. Effect
sizes, Cohen’s $d$ with correction for repeated measures (Morris & DeShon, 2002), for these differences were mid-range, .672, .408, and .447, respectively.

Results from a large-scale meta-analysis (Larson, Rottinghaus, & Borgen, 2002) of tests of the six Holland vocational personality dimension and “big-five” personality traits suggest a predictable relationship between several of the scores. Their analysis suggested a consistent positive relationship between the following:

- Artistic vocational personality trait and big-five Openness trait
- Enterprising vocational personality trait and big-five Extraversion trait
- Social vocational personality trait and big-five Extraversion trait
- Social vocational personality trait and big-five Agreeableness trait
- Investigative vocational personality trait and big-five Openness trait
- Conventional vocational personality trait and big-five Conscientious trait.

Table 2 displays the relationship between the 10-option and the two scenarios for simulated 3-option scoring of the vocational personality scales and the big-five personality trait scores as measured by the NEO-FFI. A perusal of Table 2 shows the expected pattern between vocational personality dimensions and big-five personality traits on four of the six relationships, Enterprising–Extraversion, Investigative–Openness, Social–Agreeableness, and Conventional–Conscientiousness. Using the Fisher $r$ to $z$ transformation, none of the differences in these four relationships between use of the 10-option and the simulated 3-option scoring of vocational personality trait scales approached statistical significance.

It is typical for scores on instruments measuring the six vocational personality dimensions to be interpreted as codes or types based on ranking of the responses on the dimensions (Holland, 1997). In the simulation study, three separate analyses explored whether the number of response alternatives impacted score rankings on the six scales: (a) analysis of the congruence with different response option numbers, (b) analysis using a two-factor interpretation in which the two factors are based on individual scale ranking, and (c) direct comparison of the ranks obtained using the 10-option and the two 3-option scenarios.

For the congruence analysis, three Holland codes (two-letter) were generated for each participant based on 10-option scoring and each of the 3-option scoring procedures. The two-letter code represents the two highest ranked scale scores for the participant, for example, a participant whose highest scores were on the Realistic and Investigative scales would have a code of RI. Three Iachan indices (Iachan, 1990) were then calculated to assess the congruence of the codes from the three scoring procedures.

With two-letter Holland codes, the Iachan congruence index can range from zero to six, with six indicating that the

**Table 2.** Simulation Study: Pearson Product–Moment Correlation Coefficients Between NEO Personality Trait Scores and Vocational Personality Traits With Three Scoring Conditions ($N = 77$).

|                  | N    | E    | O    | A    | C    |
|------------------|------|------|------|------|------|
| Realistic 10-option | .15  | −.03 | −.09 | .02  | −.16 |
| Realistic 3-option: 1 | .14  | −.09 | −.08 | .01  | −.14 |
| Realistic 3-option: 2 | .11  | −.05 | −.04 | .03  | −.16 |
| Investigative 10-option | .07  | −.02 | .31**| −.12 | −.02 |
| Investigative 3-option: 1 | .03  | .05  | .28* | −.09 | .00  |
| Investigative 3-option: 2 | .03  | −.04 | .31**| −.05 | .00  |
| Artistic 10-option | −.02 | .05  | .03  | −.19 | −.13 |
| Artistic 3-option: 1 | −.01 | .07  | −.01 | −.22*| −.11 |
| Artistic 3-option: 2 | .00  | .07  | −.06 | −.25*| −.15 |
| Social 10-option | .08  | −.18 | −.12 | .21  | −.17 |
| Social 3-option: 1 | .12  | −.22*| −.09 | .25* | −.20 |
| Social 3-option: 2 | .14  | −.20 | −.10 | .23* | −.17 |
| Enterprising 10-option | −.03| .35**| .17  | −.23*| .08  |
| Enterprising 3-option: 1 | −.05| .36**| .16  | −.26*| .06  |
| Enterprising 3-option: 2 | −.09| .40**| .16  | −.22 | .10  |
| Conventional 10-option | −.21| −.17 | −.24*| .28* | .34**|
| Conventional 3-option: 1 | −.19| −.19 | −.25*| .25* | .31**|
| Conventional 3-option: 2 | −.16| −.22 | −.26*| .24* | .32**|

Note: $N$ = Neuroticism; $E$ = Extraversion; $O$ = Openness; $A$ = Agreeableness; $C$ = Conscientious. 3-Option: 1 is scenario one with original scores of 5 coded as 2. 3-Option: 2 is scenario two with original scores from 4 to 6 coded as 2.

*p < .05. **p < .01.
two-letter codes are identical. Iachan index scores of five or six are interpreted as highly congruent; index scores of three and four are interpreted as moderately congruent; index scores less than three are interpreted as incongruent (Cowger, Bickham, Miller, & Springer, 1999). The mean congruence index for the codes based on the 10-option scoring and codes based on the simulated 3-option-scenario 1 scoring in this study was at the level identified as highly congruent \((M = 5.0, SD = 1.30)\). A highly congruent outcome was also evident with 10-option scoring and the 3-option-scenario 2 scoring \((M = 5.1, SD = 1.40)\) and with both simulated 3-option scoring scenarios \((M = 5.3, SD = 1.29)\).

Data were available in the simulation study that enabled the possible influence of ability on congruence of the 10-option and 3-option scoring procedures to be tested. For the independent variable, the Iachan congruence index scores were divided into two categories: high congruence, defined as indices 5 or greater, and moderate/incongruent, defined as indices less than 5. Participant scores on the WJ-III GIA served as the dependent variable.

Of the 77 participants, 62 had a high level of congruence in codes from the 10-option and 3-option-scenario 1 rankings. There was no apparent relationship between ability and extent of congruence, \(F(1, 75) = 0.230, p = .633,\) effect size (partial eta squared) = .003. Similar results were obtained for congruence between 10-option scoring and 3-option-scenario 2 scoring (66 of the 77 participants had a high level of congruence) with no statistically significant difference in scores on the ability test, \(F(1, 75) = 0.050, p = .824;\) results were also similar for congruence between the two simulation scenarios \((n = 62\) in the high congruence category) with no statistically significant difference in scores on the ability test, \(F(1, 75) = 0.077, p = .782.\) Effect sizes were .001 and .001, respectively.

A comparable procedure was used to assess the possible influence of personal distress on the utility of the three scoring procedures with the NEO Neuroticism scale serving as the dependent variable. The difference in mean Neuroticism scores between participants in high congruence and moderate/incongruent Iachan index categories using 10-option and 3-option-scenario 1 scaling was not statistically significant, \(F(1, 75) = 1.444, p = .233,\) effect size = .019. Using 10-option and 3-option-scenario 2 scoring to calculate the congruence index, differences in mean Neuroticism scores were not statistically significant, \(F(1, 75) = 1.445, p = .233,\) effect size = .019. Using the congruence index between the two simulation scenarios, the differences were also not statistically significant, \(F(1, 75) = 0.002, p = .967,\) effect size = .000.

Prediger, Swaney, and Mau (1993) proposed a two-factor solution, devising a procedure that uses a weighted ranking of the six Holland dimensions to create two summary scales. One creates a “preference for working with things” versus “preference for working with people” scale, and the other creates a scale with a continuum of preference for working with data or preference for working with ideas.

Differences in mean scores on the things-people scale for the three scoring conditions in the simulation study were not statistically significant. Using paired samples analysis, the results were as follows: 10-option versus 3-option-scenario 1, \(t(76) = 1.773, p = .080,\) and 10-option versus 3-option-scenario 2, \(t(76) = 1.395, p = .167.\) The difference between the two 3-option scenarios was also not statistically significant, \(t(76) = 0.855, p = .395.\) Effect sizes, Cohen’s \(d\) with correction for repeated measures, were .198, .159, and .093, respectively.

On the data-ideas scale, there was a statistically significant difference when comparing the mean score using 10-option scoring with the mean score using 3-option-scenario 2 scoring. Other comparisons were not statistically significant, and all effect sizes were small. The results on the data-ideas scale were 10-option versus 3-option-scenario 1, \(t(76) = 1.954, p = .054;\) 10-option versus 3-option-scenario 2, \(t(76) = 2.442, p = .017;\) 3-option-scenario 1 versus 3-option-scenario 2, \(t(76) = 0.013, p = .990.\) Effect sizes were .217, .274, and .002, respectively.

The Wilcoxon Signed Ranks Test for related samples was used for direct comparison of the ranks of the six Holland scales obtained with each of the three scoring conditions. Of the 18 comparisons, without Bonferroni correction for multiple comparisons, only 2 were statistically significant: rank of the Artistic scale in the 10-option scoring condition versus rank of the Artistic scale in the 3-option-scenario 1 simulation \((z = 2.642, p = .008)\) and rank from 10-option scoring of the Artistic scale versus rank of the Artistic scale from 3-option-scenario 2 \((z = 2.804, p = .005).\) The effect sizes (Corder & Foreman, 2009) for the two comparisons, .301 and .320, were low, and neither of the differences would reach the .05 level of significance with the Bonferroni correction.

Split-half (alternating items) reliability coefficients were calculated for each of the six scales in the 10-option scoring and the two 3-option simulation scenarios and corrected using the Spearman–Brown prophecy formula. Results are displayed in Table 3. With the Bonferroni correction for multiple comparisons, none of the differences in reliability estimates produced by the different number of response options were statistically significant.

**Study 2**

Results of the simulation study challenged the belief that additional response alternatives would have meaningful impact on psychometric characteristics of a vocational personality scale. But there was an obvious limitation with the varied number of responses simulated rather than directly tested. It may, for example, be a reasonable assumption that participants who chose a middle response in the standard format would also have chosen a middle response in a 3-point format. And, it may be reasonable to assume that those who selected among the several choices on either side of the
mid-point would have chosen the option in that direction if only one had been available. But, these, particularly regarding participants who did not choose a middle response, are only assumptions that could not be directly addressed in a simulation design.

Study 2 was a follow-up using the same core instrument but with a design enabling direct comparison of the impact of varying the number of alternative responses.

**Study 2 Participants**

Data for Study 2 were provided by 192 participants, volunteers who chose this study from several different projects available to meet a research participation requirement for courses in educational psychology. The study was conducted in the summer and fall of 2010. Most participants (79%) were female. The majority (59%) were in the age range of 18 to 24 with 28% reporting an age range of 25 to 34, 9% reporting an age range of 35 or older. Self-reported ethnic backgrounds were Caucasian (60%), Hispanic American (16%), African American (8%), Asian American (6%), and Other (9%). English was the primary language for all participants in Study 2. The range of self-reported grade point average of Study 2 participants on a 4.0 scale was 3.5 to 4.0 (41%), 3.0 to 3.4 (42%), and 2.5 to 2.9 (17%).

**Study 2 Instrumentation**

*CogStyle* was again used as the core scale. In Study 2, two versions of the 30-item scale were used. In the first version, the comparison mode used the original binary form of the instrument. Participants, in forced-choice format, expressed a preference (“more likely to be”) for one of the two adjectives or actions (e.g., efficient or curious). In the alternate version, the extent of preference was reported in a 6-point Likert format (e.g., much more likely to be efficient, somewhat more likely to be efficient, slightly more likely to be efficient, slightly more likely to be curious, somewhat more likely to be curious, much more likely to be curious).

**International Personality Item Pool–50 (IPIP-50).** In Study 2, the big-five personality characteristics were measured using the 50-item scale from the IPIP (http://ipip.ori.org/ipip/). The IPIP (Goldberg et al., 2006) is an open source resource with sample questionnaires for personality constructs including the five-factor model of Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to Experience. Reliability estimates (Cronbach’s alpha) in this participant sample for the Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to Experience scales were .88, .84, .77, .83, and .75, respectively.

**Vocational Efficacy Scale.** This subject pool study also included a measure of vocational efficacy organized around the Holland six dimensions. Participants were instructed to respond to questions about the extent to which they felt confident in their ability to perform various tasks associated with each of the dimensions. Reliability estimates, Cronbach’s alpha, for the efficacy dimensions of Realistic, Investigative, Artistic, Social, Enterprising, and Conventional were .79, .78, .74, .77, .77, and .77, respectively.

**Study 2 Procedure**

All scales were completed in an online format. Participants were randomly assigned to complete the *CogStyle* scale in binary format or the alternate version with preferences presented in a 6-point Likert format. Typical for studies in subject pool online modalities, cases in which motivation was suspect were identified and deleted before additional analysis. Suspect data were defined on the basis of overall time spent completing the several instruments. A total of 218 data sets were examined; 16 were eliminated on the basis of completion time too short to indicate other than random responding. The analysis was thus based on a total of 192 participants, 95 of whom completed the original binary form of the *CogStyle* scale and 97 who completed the Likert response form.

The random assignment appeared to identify equivalent groups with no statistically significant difference in the demographics of the two groups evident on gender,
Table 4. Study 2: Descriptive Statistics for Vocational Personality Trait Scores With Binary and 6-Point Likert-Type Response Options.

|               | Realistic | Investigative | Artistic | Social | Enterprising | Conventional |
|---------------|-----------|---------------|----------|--------|--------------|--------------|
| **M (SD)**    | **M (SD)**| **M (SD)**    | **M (SD)**| **M (SD)**| **M (SD)**    | **M (SD)**    |
| Binary (n = 95) | 4.1 (2.11) | 4.6 (2.16)    | 4.9 (1.96) | 6.7 (2.18) | 3.9 (2.25)    | 5.8 (2.63)    |
| 6-choice (n = 97) | 2.7 (1.65) | 2.9 (1.68)    | 3.2 (1.91) | 4.7 (2.10) | 2.4 (1.66)    | 4.1 (2.28)    |

Note: Statistically significant differences (p < .01) were evident between all mean scores.

Table 5. Study 2: Pearson Product–Moment Correlation Coefficients Between Big-Five Personality Trait Scores (IPIP) and Two Scoring Conditions for Vocational Personality Traits.

|               | N          | E          | O          | A          | C          |
|---------------|------------|------------|------------|------------|------------|
| R-Binary Scale| .06        | -.25**     | -.05       | -.28**     | .20        |
| R-Likert Scale| .18        | -.17       | -.24*      | -.23*      | .38***     |
| I-Binary Scale| -.09       | -.01       | .05        | -.12       | -.45**     |
| I-Likert Scale| -.12       | .04        | .21*       | -.09       | -.28**     |
| A-Binary Scale| .02        | -.06       | .23*       | .05        | -.03       |
| A-Likert Scale| -.15       | .00        | .08        | .09        | -.20*      |
| S-Binary Scale| .01        | .19        | -.17       | .37***     | -.17       |
| S-Likert Scale| .07        | .21*       | -.07       | .51**      | -.02       |
| E-Binary Scale| .02        | .27**      | .18        | -.02       | -.26*      |
| E-Likert Scale| .00        | .31**      | .37**      | .05        | -.15       |
| C-Binary Scale| -.02       | -.14       | -.19       | .00        | .59***     |
| C-Likert Scale| -.03       | -.21*      | -.21*      | -.12       | .61***     |

Note: IPIP = International Personality Item Pool. Big-Five: N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientious. Binary-Likert: R = Realistic; I = Investigative; A = Artistic; S = Social; E = Enterprising; C = Conventional.

χ²(1, 192) = .914, p = .339; age range, χ²(3, 191) = .656, p = .883; ethnicity, χ²(4, 190) = 9.149, p = .06; or grade point average, χ²(2, 184) = 1.656, p = 1.883.

The randomly assigned groups were also compared on results of the IPIP-50 scale used to test the big-five personality characteristics as a validity measure in Study 2. No statistically significant differences were found on the Neuroticism trait between the binary (M = 32.3) and Likert (M = 31.2) groups, t(190) = 0.992, p = .322; on the Extraversion trait between the binary (M = 34.5) and Likert (M = 33.1) groups, t(190) = 1.162, p = .247; on the Openness trait between the binary (M = 37.9) and Likert (M = 37.5) groups, t(190) = 0.455, p = .650; on the Agreeableness trait between the binary (M = 41.6) and Likert (M = 42.1) groups, t(190) = 0.712, p = .477; or on the Conscientiousness trait between the binary (M = 36.8) and Likert (M = 35.7) groups, t(190) = 1.053, p = .294. Effect sizes (Cohen’s d) were .143, .168, .066, .103, and .152, respectively.

Study 2 Results

Consistent with the procedure described in the simulation study, the total score for each dimension was weighted by the extent of preference. For example, the adjective “efficient” is associated with the Conventional scale in the Holland model; “curious” is associated with the Investigative scale. In the binary format, the participant receives a score of 1 on one of these two scales dependent on the selection. In the 6-point Likert format, the points are awarded based on the extent of preference (e.g., much more likely to be efficient was 1 point on the Conventional scale, somewhat more likely to be efficient was .66 point on the Conventional scale, slightly more likely to be efficient was .33 on the Conventional scale, and so forth.)

Table 4 displays the descriptive information for the binary and the 6-point scoring for each of the six vocational personality scales. As expected, the “all or nothing” response mode resulted in higher raw scores in the binary format on all scales. All mean differences were statistically significant.

Table 5 displays the relationship between the binary and 6-point scoring of the vocational personality scales and the big-five personality trait scores. A perusal of Table 5 shows the expected statistically significant positive relationship between the Artistic vocational personality trait and big-five Openness trait (binary response mode), the Enterprising vocational personality trait and big-five Extraversion trait (both response modes), the Social vocational personality trait and big-five Extraversion trait (Likert response mode), and the Social vocational personality trait and big-five Agreeableness trait (both response modes). Moreover, as
Table 6. Study 2: Pearson Product-Moment Correlation Coefficients Between Vocational Efficacy Scores and Two Scoring Conditions for Vocational Personality Traits.

|                  | R-Efficacy | I-Efficacy | A-Efficacy | S-Efficacy | E-Efficacy | C-Efficacy |
|------------------|------------|------------|------------|------------|------------|------------|
| R-Binary Scale   | .34**      | .09        | .07        | -.19       | -.17       | .15        |
| R-Likert Scale   | .33**      | -.01       | -.13       | -.25*      | -.21       | .15        |
| I-Binary Scale   | -.33**     | .02        | -.07       | -.23*      | .01        | -.33**     |
| I-Likert Scale   | -.04       | .00        | .01        | -.07       | -.01       | -.08**     |
| A-Binary Scale   | .21*       | .02        | .39***     | .05        | -.05       | -.11       |
| A-Likert Scale   | .18        | -.25       | .18        | -.09       | -.08       | -.24*      |
| S-Binary Scale   | -.11       | -.15       | -.04       | .15        | .04        | -.23**     |
| S-Likert Scale   | -.10       | -.08       | -.12       | .34**      | .01        | -.05       |
| E-Binary Scale   | -.21*      | .06        | -.16       | .15        | .29**      | -.15       |
| E-Likert Scale   | -.12       | .11        | .07        | .26*       | .32**      | -.03       |
| C-Binary Scale   | .11        | -.02       | -.11       | .06        | -.12       | .55**      |
| C-Likert Scale   | .06        | .01        | -.09       | -.06       | -.10       | .49**      |

Note: R = Realistic; I = Investigative; A = Artistic; S = Social; E = Enterprising; C = Conventional.

expected, there were statistically significant positive relationships between the Investigative vocational personality trait and big-five Openness trait (Likert mode) and the Conventional vocational personality trait and big-five Conscientious trait (both response modes).

There were only two instances where there appeared to be a marked difference contingent on the response mode in the predicted relationships with big-five personality traits, one in favor of the binary response mode and one in favor of the Likert response mode. Using the Fisher r to z transformation, none of the differences in relationships with big-five personality traits between the binary and the 6-point Likert scale approached statistical significance.

For exploration of the impact of the number of response alternatives on the rankings typically used for interpretation of scores on the six vocational personality traits, the Iachan congruence index was not appropriate in Study 2 because different participants completed the two versions of the CogStyle scale. The two-factor solution, described in the Simulation study, was used along with a direct comparison of the ranks obtained in the binary and Likert formats.

Differences in the things-people scale and data-ideas scales for the two response modes in Study 2 were not statistically significant, $t = .772, p = .441$ and $t = .24, p = .881$, respectively. Effect sizes (Cohen’s $d$) were .111 and .035, respectively.

For a direct comparison of the ranks, the nonparametric Mann–Whitney $U$ procedure was used to compare the weighted rankings obtained from binary and Likert response modes on each of the six Holland dimensions. The obtained $p$ values in comparing the rankings from the binary and Likert response modes on the Realistic, Investigative, Artistic, Social, Enterprising, and Conventional Scales were as follows: $.772, .314, .846, .286, .331$, and $.878$, respectively. Corresponding effect size estimates (Probability of Superiority) were 49%, 46%, 49%, 46%, 46%, and 49%, respectively. Grissom and Kim (2005) identifying equivalents of Cohen’s $d$ and the Probability of Superiority suggested that 56% is a small effect, 65% is a medium effect, and 71% is a large effect.

One additional scale, not available for participants in the Simulation study, was available for use as an external criteria measure in Study 2. The relationships between the binary and Likert scoring of the vocational personality traits and the scores on a vocational efficacy scale are displayed in Table 6. A perusal of Table 6 suggests that, in general, the expected relationships between vocational personality trait and vocational efficacy scores were evident with no apparent advantage for the 6-point scoring. In fact, of the statistically significant relationships, most were in favor of the binary scoring. Using the Fisher r to z transformation, however, none of the differences in relationships were statistically significant.

Split-half (alternating items) coefficients were calculated for each of the six scales in the binary and in the Likert formats and corrected using the Spearman–Brown prophecy formula. Results are displayed in Table 6. With the Bonferroni correction for multiple comparisons, none of the differences between binary and 6-point Likert scoring were statistically significant. Results with this participant sample are consistent with other reliability data for the CogStyle instrument (Jones, 2009).

Discussion of Two Studies

Cautions in generalizing to other participant samples and a need for additional study are typically found in the discussion section of published research studies but are particularly important in interpreting the results reported here. One obvious limitation in generalizing the results of these studies is the nature of the sample. All participants in both studies were university students who were participating to meet a research
considering the limitations described above, these studies are not sufficient alone for conclusions about the optimal number of response categories. These findings, however, do appear sufficient for a “that is correct” response to the suggestion in the title of this article that “more may not be better” and to support a need for additional studies comparable to Study 2 with random assignment of the number of response options, more diverse participant samples, and perhaps a core instrument that does not use ipsative scoring.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69*, 49-56.
Bendig, A. W. (1954). Reliability and the number of rating scale categories. *Journal of Applied Psychology, 38*, 38-40.
Blaikie, N. (2003). *Analysing quantitative data*. London, England: SAGE.
Bloom, M., Fischer, J., & Orme, J. G. (2003). *Evaluating practice: Guidelines for the accountable professional* (4th ed.). Boston, MA: Allyn & Bacon.
Cook, C., Heath, F., Thompson, R. L., & Thompson, B. (2001). Score reliability in web- or Internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales. *Educational and Psychological Measurement, 61*, 697-706.
Corder, G. W., & Foreman, D. I. (2009). *Nonparametric statistics for the non-statistician*. Hoboken, NJ: John Wiley.
Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Lutz, FL: Psychological Assessment Resources.
Cowger, E. L., Bickham, P. J., Miller, M. J., & Springer, T. P. (1999). Accuracy of leisure activities finder among high school students. *College Student Journal, 33*, 444-449.

Table 7. Study 2: Split-Half Reliability Estimates With Spearman–Brown Prophecy Formula for Two Scoring Conditions on Vocational Personality Traits.

| Trait       | Reliability estimate binary \( n = 95 \) | Reliability estimate Likert \( n = 97 \) |
|-------------|------------------------------------------|------------------------------------------|
| Realistic   | .72                                      | .73                                      |
| Investigative | .64                                      | .62                                      |
| Artistic    | .53                                      | .69                                      |
| Social      | .62                                      | .79                                      |
| Enterprising| .69                                      | .75                                      |
| Conventional| .77                                      | .84                                      |

requirement. The number of participants, particularly in the Simulation study, was relatively small. The Simulation study included only two of the many possible simulation scenarios for the number of response options. In addition, the extent to which a simulation will reflect an actual response choice in a real-life setting is always uncertain.

In reference to the first research question, the impact on measurement statistics, the only clearly evident effect in either of these studies was an increase in the mean scores when the binary (all or nothing) results were compared with six response categories with partial credit. While no definitive conclusions could be drawn from the split-half reliability estimates in the two studies, there was no apparent consistent enhancement or decrement of the estimated reliability with the different numbers of response options. In the Simulation study, most reliability estimates were low under all conditions, but there was no trend indicating advantage or disadvantage of the number of options. Reliability estimates in Study 2 trended higher for the Likert response options compared with the binary option, but the differences did not rise to a level of statistical significance.

Possibly confounding interpretation of the measurement statistics, including the reliability estimates, in these studies is that the forced-choice format in the core instrument results in ipsative scores, an artificial interdependence among the six scales that may have an influence on psychometric properties. While it has been suggested (Baron, 1996) that ipsative scoring may actually control for bias that is inherent in Likert-type responses, further study using a core instrument with normative scoring would be advantageous.

In reference to Research Question 2, neither study found evidence indicating increased utility in prediction of related external criteria associated with the number of response categories. In the Simulation study, the two scenarios with 3 response categories appeared as effective as 10 response categories in predicting scores on a measure of core personality traits and in creating ranks used for interpretation of vocational personality dimensions. In Study 2, 6 response categories appeared no more effective than 2 response categories in predicting scores on a different measure of core personality traits, creating ranks used for interpretation of the vocational personality dimensions, or in predicting scores on a self-report of vocational efficacy.
Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. Journal of Research in Personality, 40, 84-96.

Grissom, R. J., & Kim, J. J. (2005). Effect sizes for research: A broad practical approach. Mahwah, NJ: Lawrence Erlbaum.

Holland, J. L. (1973). Making vocational choices. Englewood Cliffs, NJ: Prentice Hall.

Holland, J. L. (1997). Making vocational choices: A theory of vocational personalities and work environments (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Iachan, R. (1990). Some extensions of the Iachan congruence index. Journal of Vocational Behavior, 36, 176-180.

Jamieson, S. (2004). Likert scales: How to abuse them. Medical Education, 38, 1217-1218.

Jones, W. P. (1996). Career assessment for patients with visual disability: Consideration of instrument distortion. Journal of Vision Rehabilitation, 10, 2-7.

Jones, W. P. (2009). CogStyle core scale reliability. Retrieved from http://faculty.unlv.edu/pjones/nl8_1.htm

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. Nursing Research, 39, 121-123.

Krieg, E. F., Jr. (1999). Biases induced by coarse measurement scales. Educational and Psychological Measurement, 59, 749-766.

Larson, L. M., Rottinghaus, P. J., & Borgen, F. H. (2002). Meta-analyses of big six interests and big five personality factors. Journal of Vocational Behavior, 61, 217-239.

Lietz, P. (2010). Research into questionnaire design: A summary of the literature. International Journal of Market Research, 52, 249-272.

Lowen, W. (1982). Dichotomies of the mind: A systems science model of the mind and personality. New York, NY: John Wiley.

Mattell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. Educational and Psychological Measurement, 31, 657-674.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological Methods, 7, 105-125.

Prediger, D., Swaney, K., & Mau, W. (1993). Extending Holland’s hexagon: Procedures, counseling applications, and research. Journal of Counseling & Development, 71, 422-428.

Rodgers, W., Andrews, F., & Herzog, R. (1992). Quality of survey measures: A structural modeling approach. Journal of Official Statistics, 8, 251-275.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson Tests of Cognitive Abilities (WJ-III). Itasca, IL: Riverside.

Author Biographies

W. Paul Jones, Ed.D. is a Professor in the Department of Educational Psychology & Higher Education at the University of Nevada, Las Vegas, where he directs doctoral studies in school psychology. His current research interests include adapting scales for online administration and single-case statistical analysis.

Scott A. Loe, Ph.D., is an Associate Professor in the Department of Educational Psychology & Higher Education at the University of Nevada, Las Vegas, where he directs clinical studies in school psychology. His current research interests include assessment training models and school wide positive behavior support.