Impact of Survey Design Features on Score Reliability

Brian K. Miller¹, ², Marcia Simmering²
¹ Department of Management, Texas State University, TX, US, ² Department of Management, Louisiana Tech University, LA, US

Keywords: experiment, coefficient alpha, survey design, score reliability

Impact of Survey Design Features on Score Reliability

Much has been written on the psychology of survey responses (e.g., Tourangeau et al., 2000) and the science of self-report (e.g., Stone et al., 2000). Researchers often depend upon data from inventories and scales that measure psychological constructs, many of which are best accessed through self-report (Franke, 1997). Self-reports are critical because of the internalized information about which the self is best suited to respond (Paulhus & Vazire, 2007), but the veracity of the data collected directly from respondents can be suspect. In statistical terms, this veracity is one of the many aspects of an ongoing program of collecting evidence of construct validity (Benson, 1998).

It is well-known that reliability is a necessary but not sufficient condition for validity (Nunnally & Bernstein, 1994) and that the reliability of scores on tests is a characteristic of the sample at a particular administration and not a characteristic of the test itself (Thompson, 1994). Thus, reliability is at the very foundation of validity and quite appropriately, various efforts to improve the reliability of scores on instruments and scales abound. Unreliability can render as imprecise the factor structure of measures (Huang et al., 2012; Woods, 2006) and attenuate focal variable relationships (Cohen et al., 2005). Imprecise factor structures and weakened variable relationships have consistently frustrated survey researchers relying on self-reports of traits, attitudes, beliefs, values, etc. as well as clinicians and practitioners making important decisions based upon scores on self-report inventories.

This study seeks to examine the impact of study design and administration characteristics that can be manipulated before data are collected in the hopes of maximizing score reliability. Three independent variables are manipulated: (1) presence versus absence of quality control items that are sometimes referred to as attention checks, (2) grouping items from a scale next to each other in the survey versus randomly scrambling items from different scales at different places in the survey, and (3) collecting self-report data in an anonymous manner by which the respondent cannot be matched to their responses versus collecting data confidentially whereby respondents affix their name to their survey and their responses can be matched to them. This two-by-two-by-two completely crossed cluster randomized experiment (Campbell & Stanley, 1965) makes use of random assignment to treatment conditions and the dependent variable is measured as Cronbach’s coefficient alpha of internal consistency reliability with the standard deviation of alpha calculated using Hakstian and Whalen’s (1976) formulae for three different commonly used self-report scales.

Internal Consistency Reliability

According to classical test theory (CTT) the reliability of scores on a questionnaire represents the dependability, stability, or consistency of such scores (Nunnally & Bernstein, 1994). With self-report measures that use multiple items with numeric, Likert, Likert-type, etc. responses, reliability is usually measured using Cronbach’s (1951) coefficient alpha of internal consistency reliability. Notably, alpha provides a measure of internal consistency (i.e., the degree of relatedness among items) and not necessarily homogeneity or unidimensionality (Cortina, 1993). Rangeing from zero to one, alpha scores above .70 are generally considered ac-
ceptable for exploratory research but scores should be above .90 for clinical diagnoses particularly with regard to mental and physical illness (Nunnally & Bernstein, 1994). Alpha is the average correlation between all possible split-halves of a multi-item scale across respondents (Cronbach, 1951) and may vary from group to group or even from administration to administration for the same group (Thompson, 1994). Thus, reliability is not an aspect of a test, but rather it characterizes scores on a test at a given point in time for a given group of respondents. Cronbach’s alpha is only an approximation of individual-level reliability scores on an instrument in a sample because alpha is technically a measure of group-level reliability (Raju et al., 2007) and individual alphas are not mathematically possible to calculate for individual respondents in a group. In CTT, the group-level score reliability is the ratio of true score variance to observed score variance for all subjects in a sample. It is common for researchers to (falsely) assume that all members of a group provide similarly reliable responses to a unidimensional scale, but this is not possible to statistically ascertain.

Influence of Overly Random Responding on Reliability

One influence on score reliability is when test-takers answer in a random manner and their within-person responses are correlated at a near-zero level. In order to offset this, researchers sometimes use validity check items that ask respondents to indicate whether they responded truthfully and completely (e.g., Costa & McCrae, 1992). Failure to affirmatively endorse this question suggests to the researcher that other responses collected from that respondent are suspect, potentially unreliable, and therefore invalid. Other researchers include randomly inserted items into their surveys that require a specific and exact responses to non-substantive questions and can be thought of as reliability checks or quality control items. Quality control items have been empirically studied under headings such as careless responding (Kam & Meyer, 2015; Maniaci & Rogge, 2014; Meade & Craig, 2012; Meyer et al., 2013; Schmitt & Stuits, 1985; Ward & Pond, 2015), inconsistent responding (Akbulut, 2015), random responding (Beach, 1989; Credé, 2010; Lopez & Charter, 2001; Morey & Hopwood, 2004), response effort (Lim et al., 1998; Toro-Zambrana et al., 1999), and content non-responsivity (Nichols et al., 1989). However, to a respondent keenly intent on truly random responding, even quality control items may not be salient. Quality control items can be worded quite differently from study to study and imply to respondents that their responses matter and that some responses can be discovered to be incorrect. An example of such an item is: “For quality control purposes please choose ‘disagree’ here.” Thus,

Hypothesis 1: Data gathered from surveys containing quality control items will result in higher reliability than data collected without quality control items.

Another influence on score reliability is whether the data are collected in anonymous or confidential testing situations (Meade & Craig, 2012; Ward & Pond, 2015). In the confidential setting, respondents affix their name to their survey thereby voiding the anonymous nature of a data collection. Such situations are more common in employment testing and psychological/psychiatric testing where it is highly critical to match responses to respondent and, perhaps equally so, because of the purposes of these tests, the scores on such tests must be highly reliable. In many research-based data collections, individual differences are measured via anonymous responses under the premise of desiring completely truthful responses from respondents thereby minimizing the impact of socially desirable responding, amongst other objectives. In such settings, the "reduced accountability increases the probability of...random responding" (p. 109: Johnson, 2005). Anonymous data collections also likely give rise to unreliability because there is no way to trace responses back to the responder. If submitting an anonymous survey, the respondent need not expend any effort at all and is more likely to provide unreliable responses. On the other hand, submitting a survey affixed with one’s name is likely to induce effort and result in more reliable scale scores. Thus,

Hypothesis 2: Data gathered from confidential surveys will result in higher reliability than data collected anonymously.

Influence of Overly Consistent Responding on Reliability

The opposite of overly random responding is overly consistent responding, whereby respondents provide identical responses to all items measuring a particular construct. When this happens, the coefficient alpha for a group for scores on the instrument approaches unity (i.e. 1) as long as there is some between-person variability in responses. The response tendency to provide nearly identical survey responses to items designed to measure a common construct is referred to as the consistency motif (Podsakoff & Organ, 1986) and represents a natural human tendency to be perceived of as consistent and stable. This tendency is more pronounced for those respondents who perceive that similarly worded test items are designed to catch them in specific inconsistencies (Schriesheim & DeNisi, 1980) and is more likely to arise in confidential settings in which responses can be matched to respondents. An extreme form of overly consistent responding is known as "straightlining" whereby a survey respondent provides the exact same response down a straight line of items on a survey (Desimone et al., 2018). Straightlining is not an unconscious human tendency like the consistency motif nor is it an effort at avoiding detection for inconsistencies but rather it is form of insufficient effort responding. Of note is the difference between the consistency motif and acquiescence bias. With the latter, respondents engage in yea-saying or nay-saying and researchers typically use reverse scored items to discourage choosing all high or all low responses to the items in an instrument. Reverse scored items usually help alleviate the tendency of some respondents to engage in straitlining as well.

To offset the consistency motif, researchers sometimes scramble the order of the substantive items in their multi-dimensional or multi-construct instruments with the added advantage of disguising the purpose of the instrument (Franke, 1997). By scrambling items randomly in a survey, the researcher hopes that respondents will evaluate each item on its own rather than based upon the similarity of each item to nearby items measuring the same construct. Indeed, Weijters, Geuens, and Schillewaert (2009) found that positioning items from the same construct at least six items apart from one another reduced correlations among items, as compared to when items were grouped by construct. However, scrambled items can challenge respondents’ limited information-processing skills (Rush et al., 1981) by forcing them to shift back and forth between different issues and thereby heightening the intellectual de-
mands placed on them (Solomon & Kopelman, 1984). The benefit of using grouped items is that reliability for scores is usually higher than when items are scrambled (Franke, 1997; Melnick, 1995; Schriesheim et al., 1989). To examine this, Schell & Oswald (2015), conducted an *eyeball* comparison of alpha for a sample and found no difference between grouped and scrambled items and state that "... item order does not affect the underlying measurement properties of psychological instruments" (p. 320). The current study uses *statistical* comparisons of score reliability for experimental manipulations, thus,

**Hypothesis 3:** Data gathered using grouped items will result in higher reliability than data collected using scrambled items.

### Method

#### Participants

Four hundred forty-three participants enrolled in courses at a large university in the American southwest were solicited for their voluntary participation in exchange for a modest amount of extra credit in their class. The data came from one course in two very large lecture sections taught by the same professor. The in-class survey was announced two weeks ahead of time as well as a one-week reminder by the professor that extra credit could be earned by participating. An alternative extra credit assignment was offered that required students to find an article published by their professor that extra credit could be earned by participating. For each treatment group in this study, alpha is calculated using the CTT formula (Cronbach, 1951) and transformed for non-normality as in Equation 1. Because the experimental data were analyzed with analysis of variance (ANOVA), the mean alpha for each experimental condition also required a standard deviation of alpha which was approximated with Hastkian and Whalen's (1976) Equation 2 below.

\[
T_i = (1 - r_{di})^{1/3} \quad (1)
\]

\[
v_i = 18J(J - 1)(1 - T_i)^{2/3}
\]

Where, \( r_{di} \) = alpha for each group of respondents, \( J = \) number of items in the scale, \( n_i = \) sample size of each group

Then the standard deviation is calculated as the square root of \( v_i \). Recall that alpha is a single value for an entire group. Given that uni-dimensionality is a key assumption of Cronbach’s alpha, all data were first submitted to exploratory factor analysis (EFA) using principal axis factoring and a varimax rotation.

#### Measures

All self-report inventories used the same seven-point Likert response scale anchored by 1 = strongly disagree and 7 = strongly agree, which allowed the items to be scrambled on half of the surveys. The instruments described below were the only scales in the survey and were selected because they were the only scales in the survey and were selected because they were the most comprehensive in the areas of interest.
Table 1: Alpha means, standard deviations, sample sizes for experimental treatment groups on three personality scales

<table>
<thead>
<tr>
<th>Experimental factors</th>
<th>Conscientiousness</th>
<th>Entitlement</th>
<th>Work ethic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>1^a 2^b 3^c</td>
<td>Group #</td>
<td>n</td>
<td></td>
</tr>
<tr>
<td>0 0 0</td>
<td>1</td>
<td>54</td>
<td>.825</td>
</tr>
<tr>
<td>0 0 1</td>
<td>2</td>
<td>54</td>
<td>.812</td>
</tr>
<tr>
<td>0 1 0</td>
<td>3</td>
<td>53</td>
<td>.798</td>
</tr>
<tr>
<td>0 1 1</td>
<td>4</td>
<td>54</td>
<td>.798</td>
</tr>
<tr>
<td>1 0 0</td>
<td>5</td>
<td>56</td>
<td>.774</td>
</tr>
<tr>
<td>1 0 1</td>
<td>6</td>
<td>54</td>
<td>.823</td>
</tr>
<tr>
<td>1 1 0</td>
<td>7</td>
<td>54</td>
<td>.729</td>
</tr>
<tr>
<td>1 1 1</td>
<td>8</td>
<td>56</td>
<td>.719</td>
</tr>
</tbody>
</table>

Note. M = mean alpha, SD = standard deviation, n = sample size.

^a Coded as 0 = absence of quality control items, 1 = presence of quality control items
^b Coded as 0 = items from each scale randomly scrambled, 1 = items measuring same scale grouped together
^c Coded as 0 = data collected anonymously, 1 = data collected confidentially

Results

As seen in Table 1, the range of alphas in each of the experimentally created sub-samples was .719 to .825 for conscientiousness, .769 to .863 for entitlement, and .836 to .910 for work ethic. The mean alpha for the entire sample for these three scales was .79, .82, and .87, respectively. Because the three treatment variables were orthogonally implemented, these independent (i.e. manipulated treatment) variables were non-significantly correlated at .00, .00, and (because of minor differences in the cell sizes) .05. Scale scores on the three personality instruments were correlated at .33 for the relationship between conscientiousness and work ethic, at .41 for conscientiousness and entitlement, and at .25 for entitlement and work ethic (all at p < .001). These correlational values provide evidence of discriminant validity for scores on the scales thus suggesting that the respondents perceived that the items measured independent but related constructs. The correlation between the alpha for the scales could not be statistically determined because there exists only one alpha for each scale for the entire sample. Independent ANOVA tests were conducted for alpha on conscientiousness, entitlement, and work ethic for this factor were all non-significant. There were no differences in reliability on any instrument whether quality control items were present or absent in the survey and no support was found for the first hypothesis.

Hypotheses Tests

Hypothesis one stated that the reliability of scores would be higher in surveys that use quality control items than surveys that did not. The F-scores for alpha on conscientiousness, entitlement, and work ethic for this factor were all non-significant. There were no differences in reliability on any instrument whether quality control items were present or absent in the survey and no support was found for the first hypothesis.

Hypothesis two stated that the reliability of scores would be higher in surveys where data are collected confidentially than in surveys where data are collected anonymously. Al-
pha in the anonymous condition for scores on the conscientiousness scale, entitlement scale, and work ethic scale was .808, .786, and .881, respectively. In the confidential condition the alpha was .761, .828, and .856 for these scales, respectively. The $F$-scores for alpha on conscientiousness, entitlement, and work ethic for this factor were all statistical significant. Despite these significant main effects, only one of the three scales was the difference in alpha in the anonymous condition for scores on conscientiousness scale did not support the third hypothesis.

Exploratory Analysis of Interaction Effects

Because the creation of eight independent groups of respondents was made possible by the two-by-two-by-two experimental design, we engaged in some exploration of the three different two-way interaction effects and the single three-way interaction. This was accomplished with a software program written in R code. Regarding the interaction between quality control items and a confidential administration, the $F$-scores for alpha on conscientiousness, entitlement, and work ethic for this factor were all significant. However, on all three instruments the estimated marginal mean for the condition in which both quality control items were present and the data were collected confidentially was not the largest of the four marginal means. Despite the significant $F$-scores there was no support for this interaction.

For the interaction between quality control items and grouped items, the $F$-scores for alpha on conscientiousness, entitlement, and work ethic for this factor were all significant. Again, on all three instruments the estimated marginal mean for the condition in which both quality control items were present and items measuring a particular scale were grouped together was not the largest of the marginal mean for the condition in which both quality control items were present and items measuring a particular scale were grouped together.
means. Therefore, despite the significant \(F\)-scores there was no support for this interaction.

Regarding the interaction between grouped items and a confidential survey administration, the \(F\)-scores for alpha on conscientiousness and work ethic for this factor were both significant but for entitlement it was not. However, on the conscientiousness and work ethic instruments the estimated marginal mean for this interaction was less than one or more of the other marginal means. Thus, there was no support for this interaction.

For the three-way interaction between grouped items, embedded quality control items, and a confidential administration, the \(F\)-scores for alpha on conscientiousness and entitlement were significant but on the work ethic scale it was not. On neither of the personality scales for which the \(F\)-test was significant was the alpha in the three-way interaction condition larger than the other seven three-way interactions. Despite the significance of two of the effects, there was no support for the three-way interaction.

**Effect Sizes**

The three ANOVAs resulted in effect sizes measured as partial eta squared. When two groups are compared as in the main effect hypothesis tests here, partial eta squared is comparable to the squared point biserial \(r\) (Grisson & Kim, 2005) and interpreted similarly. The effect sizes for the use of quality control items ranged from only .001 to .004 with a mean of .002. Using Cohen's (1988) general guidelines for effect sizes, this loosely translates to a very small effect. The statistical power computed post hoc using G*Power software (Faul et al., 2009) to detect the mean effect was only .05, however. The effects for the confidential administration of the survey were .051 for work ethic, .140 for entitlement, and .181 for conscientiousness with a mean of .124. Power for this mean effect was .73 and the effect size can be loosely construed as a small effect (Cohen, 1988). The effect sizes for grouped survey items were .184 for conscientiousness, .051 for entitlement, and .079 for work ethic with a mean of .105. The statistical power to detect this mean effect was .59 and the effect size can be loosely interpreted again as small (Cohen, 1988). Not all effects were in the direction which was hypothesized.

**Discussion**

This study expanded upon previous research on survey design features used to bolster score reliability and compared measures of internal consistency reliability in each of three treatment conditions for scores on three different self-report inventories. To our knowledge, this is the first study to simultaneously examine the impact of the manipulated variables of presence or absence of quality control items, grouped versus scrambled items, and confidential versus anonymous data collections. The dependent variables were measured using CTT to calculate the alpha reliability of scores for groups. Group-level summary data (mean, standard deviation, and sample size) were submitted to analysis of variance tests. Previous research on these sorts of survey conditions has relied on non-statistical comparisons of group alpha in treatment conditions.

The first hypothesis about the expected improvement to reliability with the inclusion of quality control items yielded the overall weakest effect as there was no impact on score reliability on any of the three instruments. A more detailed look at the raw data indicates that, of the 220 participants in the quality control item condition, nearly all participants (99.5%) answered the three quality control items correctly. These frequencies of responses support the notion that participants were generally attentive in their answers. Nevertheless, the reliability of their scores was no better than those whose surveys did not have quality control items.

The second hypothesis about the expected improvement to reliability if surveys were administered confidentially yielded the largest average effect size across the three instruments. The premise was that having to put one's name on a survey would engender greater effort and therefore more reliable data because the survey results could be matched to the respondent. However, only for entitlement was the effect in the hypothesized direction suggesting that confidentiality may have a negative impact on score reliability.

The third hypothesis about the expected improvement to reliability when similar survey items were grouped together resulted in very similar effect size as in the confidential administration treatment. The premise for grouping items was that higher reliability would result when respondents could easily see and compare their responses to similar items to make sure they were consistent. Contrary to expectations, grouped items more often than not decreased...
the reliability of scores.

The findings of the current study indicate several areas of future research that are likely to be fruitful. First, the results of this study indicate that quality control items made no difference in the reliability of scores on these personality scales. In other words, this effort to improve reliability on these scales in this sample did not do so. Because the data indicate high attentiveness to quality control items, it may be that participants were not in need of such reminders. However, student samples may be more prone to demand characteristics that increase response effort, and research examining the usefulness of quality control items in other samples is needed. Similarly, the mixed findings regarding the reliability of scores on scales in the confidential versus anonymous data collections should be investigated in other samples and in other survey media (i.e., online data collection).

The most surprising results in this study are that when grouping like items together instead of scrambling scale items with other items in a survey, researchers can sometimes reduce the reliability of scores. This indicates that this element of survey design is more complex than anticipated. As noted previously, there are both benefits and drawbacks in terms of data quality when items are grouped together on a survey, but the degree to which this is truly effective is still unknown. While researchers are interested in increasing scale reliability with consistent responses that should be more likely to be produced when items are grouped, this grouping could create consistency that is artifactual rather than true. When items are grouped the reliability of scores may suffer because some respondents engage in some purposeful deviations from consistency so as not to be seen as straightlining thereby providing downward pressure on internal consistency reliability. That is, they may occasionally provide a slightly different response to an item in a set of like items so as not to be caught in, or accused of, straightlining. On the other hand when items are scrambled straightlining is not likely as it would require a search for similar items randomly placed in a survey. Thus, a more truthful response may be provided to those items that is not dependent on remembering one's responses to other items measuring the same construct placed elsewhere in the survey. In sum, recall is not necessary when responding consistently truthfully to scrambled items from a particular scale thus applying upward pressure to reliability. When lying or faking becomes the task, remembering one's previous lies is a problem and the responses are likely to become somewhat random when not truthful. Indeed, many researchers are leery of the possibility of scores that are inflated by common method variance, which can occur when several self-report measures are included in the same survey. In this circumstance, respondents are more likely to be able to produce consistent answers to items because they have perfect recall of adjacent items previously answered, and such consistency may actually be an artifactual response bias rather than a truly reliable measure of the substantive scale (Podsakoff et al., 2003).

To explore the role of item placement further in the current data, correlations among the overall scale scores were examined in the grouped versus scrambled conditions. The bootstrapped confidence intervals for the three scale score correlations overlapped for both conditions, indicating no significant difference whether the items were grouped or scrambled. It is notable that the correlation between entitlement and work ethic was non-significant in both conditions. The lack of difference between the conditions among these relationships suggests either that (a) respondents were adept at changing their focus in order to provide similar responses to items from the same scale whether grouped or scrambled, or (b) the nature of the items was so similar that it did not matter that they were scrambled or grouped. Both of these suggests a potentially weak manipulation. On the other hand, the non-significant correlation between entitlement scores and work ethic scores in both conditions suggests ample discriminant validity exists between scores on the scales. See Table 5 for these correlations as well as those in the other main effect manipulations. In the current data, despite finding that reliability sometimes tends to be poorer in surveys with grouped items, it appears that there is no consistent finding regarding the impact of item grouping on the convergent validity of the variables. Thus, future research should continue to examine the degree to which survey item order affects score reliability.

These caveats are in line with recommended constraints on the generality of findings in the social sciences (Simons, Shoda, & Lindsay, 2017) and “we have no reason to believe that the results depend on other characteristics of the participants, materials, or context” (pp. 1126-1127). The target population is adult survey respondents in general and our sample was comprised of undergraduate students at one university which limits the generalizability of the results. The sample size also limits generalizability in that the statistical power to detect an effect where one truly exists fell short of the standard of .80. Regarding the instruments being used here, it should be noted that they may have suffered from multidimensionality and it is well-known that Cronbach’s alpha is best suited for unidimensional measures. Our exploratory factor analysis revealed that a two-factor solution fit the conscientiousness data best, either a one or two-factor solution depending on whether one gives more credence to eigen values or to the scree plot fit the work ethic data best, and a one-factor solution was supported for the entitlement data regardless of whether interpreting the eigen values or the scree plot. The procedure was quite common in that paper and pencil surveys were administered in both confidential and anonymous situations mimicking the real-world implementation. However, given this sample’s likely familiarity with surveys frequently administered for exchange in extra credit in the college classroom, the results could be affected by both demand characteristics and previous experience with being allowed to engage in random responding with no consequence.

While not a limitation of the current study per se, researchers are leery about the nature of coefficient alpha when the conditions are warrant. As noted previously, alpha measures internal consistency, and thus, one cannot assume unidimensionality of items on a scale. Further, alpha increases as the number of scale items increases, even when more than one factor underlies a scale (Pedhazur & Schmelkin, 2013). Thus, scores on alpha can be misleading. However, as coefficient alpha often represents a lower bound of reliability, it remains an important metric in social science research (Cortina, 1993) and provides a foundation for validity.

This study provides meaningful implications for researchers. First, the use of quality control items is not likely to be beneficial in improving reliability, particularly as more experienced groups of survey respondents (e.g., Amazon.com Mechanical Turk workers) gain familiarity with such items. Second, whether collecting data anonymously or confidentially seems not to matter. Third, researchers should balance concerns about score reliability, construct reliability, construct validity, and common method variance when determining whether or not to group or scramble construct items in surveys, as the current survey indicates that...
Table 5: Scale score correlations in each treatment condition

<table>
<thead>
<tr>
<th>Presence of quality control items</th>
<th>Absence of quality control items</th>
<th>Items randomly scrambled</th>
<th>Items grouped together</th>
<th>Confidential data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Conscientiousness</td>
<td>--</td>
<td>-.445 (-.557, -.327)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2. Entitlement</td>
<td>.337 (.208, .447)</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3. Work ethic</td>
<td>.337 (.208, .447)</td>
<td>-.202 (-.317, -.066)</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Presence of quality control items</th>
<th>Items randomly scrambled</th>
<th>Items grouped together</th>
<th>Confidential data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Conscientiousness</td>
<td></td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>2. Entitlement</td>
<td>-.463 (-.652, -.241)</td>
<td>--</td>
<td>-.463 (-.652, -.241)</td>
</tr>
<tr>
<td>3. Work ethic</td>
<td>.331 (0.65, 536)</td>
<td>.004 (.262, .268)</td>
<td>.331 (0.65, 536)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Presence of quality control items</th>
<th>Anonymous data collection</th>
<th>Confidential data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Conscientiousness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Entitlement</td>
<td>-.463 (-.646, -.254)</td>
<td></td>
</tr>
<tr>
<td>3. Work ethic</td>
<td>.331 (0.662, 534)</td>
<td>.004 (.266, .284)</td>
</tr>
</tbody>
</table>

Note. Bootstrapped 95% confidence intervals next to correlations in parentheses. Correlations above and below the diagonal are for different levels of each manipulation.

this survey design decision is complex. As noted by McGrath (1981), each methodological decision made by a researcher has consequences that affect other elements of the study. In conclusion, the many machinations which survey designers go to in order to improve the reliability of scores likely has a minimal impact.

Contributions

Contributed to conception and design: BKM
Contributed to acquisition of data: BKM
Contributed to analysis and interpretation of data: BKM, MS
Drafted and revised the article: BKM, MS
Approved the submitted version for publication: BKM, MS

Acknowledgements

Thanks go to David Rindskopf for a custom R program that runs factorial ANOVA based solely on summary data. This program allowed us to explore interaction effects in our experiment.

Funding Information

The authors received no funding for this research.

Competing Interests

No competing interests exist.

Data Accessibility Statement

The online Texas Data Repository (dataverse.tdl.org/dataverse/txstate) is used to share datasets through the Texas Digital Library and managed by local Texas State University librarians. The Texas Digital Library (TDL) is a consortium of academic libraries in Texas with a proven history of providing shared technology services to support secure, reliable access to digital collections of research and scholarship. The Texas Data Repository is a project of the TDL and its member institutions to develop a consortial statewide research data repository for researchers at Texas institutions of higher learning. Data is curated in the repository following accepted standards (NISO Framework Advisory Group, 2007). The persistent identifier, a DOI, used for the data, survey items, and code book in this study is https://doi.org/10.18738/T8/Y3HT9K

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.
REFERENCES


SUPPLEMENTARY MATERIALS

Peer Review Comments