UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Miller, B. K., & Simmering, M. (2020). Impact of Survey Design Features on Score Reliability. *Collabra: Psychology*, *6*(1). https://doi.org/10.1525/collabra.17975

Methodology and Research Practice

Impact of Survey Design Features on Score Reliability

Brian K. Miller[1] [a], Marcia Simmering[2]

[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

https://doi.org/10.1525/collabra.17975

Collabra: Psychology

Vol. 6, Issue 1, 2020

The a priori impact of survey design and implementation tactics on score reliability is not well-understood. Using a two-by-two-by-two cluster randomized post-test only experimental design, the Cronbach's coefficient alpha of internal consistency reliability of scores on three personality scales is calculated. The experimental conditions are presence versus absence of quality control items, anonymous versus confidential administration conditions, and randomly scrambled items versus grouped survey items. Alpha was calculated for each of the eight treatment groups. Hakstian and Whalen's (1976) formulae were used to calculate the standard deviation of alpha. These summary data were then used in analysis of variance tests. The ANOVA results were mixed for the three personality scales. The use of quality control items had no impact on alpha on any scale, confidentiality improved alpha on one scale and decreased it on two others, and grouping items together improved alpha on two scales and decreased it on another. Although most of the exploratory interaction tests for each scale were statistically significant, none were in the direction implied by the confluence of main effect hypotheses. These mixed results suggest that a priori machinations by survey designers and administrators may often result in unwanted differences in 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., 2003). 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, 1963) 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). Ranging from zero to one, alpha scores above .70 are generally considered ac-

a bkmiller@txstate.edu

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 multidimensional 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 (Soloman & Kopelman, 1984). The benefit of using grouped items is that reliability for scores is usually higher than when items are scrambled (Franke, 1997; Melnick, 1993; Schriesheim et al., 1989). To examine this, Schell & Oswald (2013), 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 and type the abstract and send it as an email to the professor. No student chose that option. The data collection was determined to be exempt by the university IRB with approval number 2658. Complete data on all survey items was provided by 435 respondents. Eight failed to complete the survey. The raw data, survey items, and code book are accessible at https://doi.org/10.18738/T8/Y3HT9K . The mean age was 22.23 years and 60% were male. Self-reported racial and ethnic group membership was 72.8% white, 5.4% black, 17% Hispanic, 2.3% Asian, .7% Native American Indian, and 1.8% "other". Almost two thirds (63.6%) were currently employed and of those, 19.4% held full-time jobs. Of the currently employed participants, 20.1% were managers of other employees, and the mean number of direct report subordinates was 10.88.

Manipulations

Treatment assignments were determined by simple random assignment. This intervention was unknown to participants in the experiment until debriefing afterward. To enact two of the three treatments, four differently colored surveys with manipulations for presence or absence of quality control items and scrambled versus grouped substantive survey items were distributed in one of two large sections of a required college course. The third treatment resulted from a coin flip to determine the cluster assignment regarding which class received the anonymous surveys and which received the confidential surveys. The paper-and-pencil surveys required each respondent to write, by hand, their numeric response in a blank to the right of each item. Because of slightly different numbers of participants in each treatment condition the design is an unbalanced experiment.

Presence or absence of quality control items. The first manipulated factor was the presence or absence of quality control items. Randomly inserted into half of the surveys were the following three items: "For quality control purposes, please choose disagree here," "For quality control purposes, please write the number three here," and "For quality control purposes, choose somewhat agree here." One quality control item was embedded at the one quarter point into the survey, at the halfway point into the survey, and at the three-fourths point in the survey. In the other half of the surveys, these items were absent.

Scrambled versus grouped survey items. The second manipulated factor was regarding the sequencing of the survey items. Half of the surveys had all items measuring each trait grouped together. The other half had the items measuring the traits randomly scrambled with each other.

Anonymous versus confidential data collection. The third manipulated factor was whether the data were collected anonymously or confidentially. In the anonymous condition, respondents omitted their name and any identifying information from their survey. In the confidential condition, respondents wrote their name on their survey and were told that their data could be tracked to them so that confidential survey feedback could be individually provided to them.

Calculation of Dependent Variables and Standard Deviations

In CTT, alpha is an approximation of the average reliability for scores for the group and is therefore a single value for the group and not calculatable for any one individual respondent. 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 Hakstian and Whalen's (1976) Equation 2 below.

$$T_i = (1 - r_{ai})^{1/3} \quad (1)$$

$$v_i = \frac{18J_i(n_i - 1)(1 - T_i)^{2/3}}{(J_i - 1)(9n_i - 11)^2} \quad (2)$$

Where, r_{ai} = alpha for each group of respondents,
J_i= 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 of their likely discriminant validity. Yet, they were not so dissimilar as to influence the effects induced by the scrambling versus grouping manipulation.

Conscientiousness. The first scale measured conscientiousness using the ten-item scale from Goldberg's (1999) International Personality Item Pool (IPIP). Sample items include: "I pay attention to details" and "I leave my belongings around" (reverse scored). Exploratory factor analysis with an oblique rotation revealed two factors with Eigenvalues greater than 1. The first three eigen values were 3.56,

Table 1: Alpha means, standard deviations, sample sizes for experimental treatment groups on three personality scales

Coefficient alpha, SD of alpha, and cell size in each experimental condition										
Experimental factors					Conscientiousness		Entitlement		Work ethic	
1[a]	2[b]	3[c]	Group #	*n*	M	SD	M	SD	M	SD
0	0	0	1	54	.825	.052	.773	.051	.836	.053
0	0	1	2	54	.812	.052	.769	.051	.882	.055
0	1	0	3	53	.798	.052	.782	.052	.895	.056
0	1	1	4	54	.798	.051	.820	.053	.910	.056
1	0	0	5	56	.774	.049	.801	.051	.845	.052
1	0	1	6	54	.823	.052	.836	.053	.848	.053
1	1	0	7	54	.729	.048	.863	.055	.890	.055
1	1	1	8	56	.719	.047	.810	.051	.841	.052

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

1.28, and .93. The scree plot also indicated that two factors best represented the variance in the scale items. This is likely the result of several reverse-score items in the scale (Schmitt & Stuits, 1985). Cronbach's coefficient alpha of internal consistency reliability for the whole sample was .79 with slightly different values in the eight treatment conditions.

Entitlement. The second scale measured trait entitlement with the eight items recommended by B. K. Miller (2009) from the instrument by Sauley & Bedeian (2000). Entitlement is a trait that predisposes one toward a preference for more than others regardless of one's effort, contribution, or performance (B. K. Miller, 2009). Sample items include: "When I am at my job I think of ways to get out of work" and "It is really satisfying to me when I can get something for nothing at work." The EFA as well as the scree plot both suggested a 1-factor solution for these responses. The first two eigen values were 3.55 and .85. Cronbach's alpha for this entire sample was .82 with slightly different values in the eight treatment conditions or subsamples.

Work ethic. The third scale measured work ethic using the ten-item Hard Work sub-scale from the Multidimensional Work Ethic Profile (MWEP) from M. J. Miller et al. (2002). Work ethic is a trait that predisposes one toward hard work often for the sake of hard work alone and the belief that there is value in hard work regardless of outcome. Sample items include: "If you work hard you will succeed" and "Hard work makes one a better person." The EFA resulted in two factors with Eigenvalues greater than 1 but the scree plot suggested that a one-factor solution explained the majority of the variance in the items. The first three eigen values were 4.72, 1.06, and .89. Alpha for the complete sample of all participants was .87 with slightly different values in the eight treatment conditions.

See Table 1 for the alpha values, their standard deviations, and sample sizes for each scale below in each of the eight treatment conditions (i.e. sub-samples created experimentally).

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 entitlement. 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) .03. 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 -.23 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 each instrument instead of a multivariate analysis of variance test. Thus, there are three different tests of each hypothesis. See Tables 2, 3, and 4 for these results.

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-

Table 2: Analysis of variance results for alpha for scores on conscientiousness scale in 2x2x2 experiment

Factors	Type III sum of squares	*df*	Mean Square	F-score	*p*-value	Partial eta-squared
Model	268.268	8	33.534	13193.313	.000	.996
Quality control items (Q)	.005	1	.005	1.807	.180	.004
Confidential data (C)	.240	1	.240	94.484	.000	.181
Grouped items (G)	.245	1	.245	96.505	.000	.184
Q x C	.018	1	.018	7.228	.007	.017
G x Q	.014	1	.014	5.657	.018	.013
G x C	.079	1	.079	31.181	.000	.068
Q x C x G	.035	1	.035	13.858	.000	.031
Error	1.085	427	.003			
Total	269.353	435				

Table 3: Analysis of variance results for alpha for scores on entitlement scale in 2x2x2 experiment

Factors	Type III sum of squares	df	Mean Square	F-score	p-value	Partial eta-squared
Model	283.550	8	35.444	13105.994	.000	.996
Quality control items (Q)	.002	1	.002	.643	.423	.002
Confidential data (C)	.187	1	.187	69.233	.000	.140
Grouped items (G)	.063	1	.063	23.155	.000	.051
Q x C	.018	1	.018	6.794	.009	.016
G x Q	.014	1	.014	5.316	.022	.012
G x C	.004	1	.004	1.447	.230	.003
Q x C x G	.115	1	.115	42.460	.000	.090
Error	1.155	427	.003			
Total	284.704	435				

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 statistically significant. Despite these significant main effects, on only one of the three scales was the difference in alpha in the direction of that which was hypothesized. Thus, there was only partial support for the second hypothesis.

Hypothesis three stated that the reliability of scores would be higher in surveys where items from a scale are grouped together than in surveys in which items are scrambled with items from other scales. The mean alpha in the grouped condition for scores on the conscientiousness scale, entitlement scale, and work ethic scale was .761, .819, and .884, respectively. In the scrambled condition the alpha was .808, .795, and .853 for these scales, respectively. The *F*-scores for alpha on conscientiousness, entitlement, and work ethic for this factor were all significant. Despite these significant main effects, only two of the three differences were in the hypothesized direction Thus, there was only partial support for 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

Table 4: Analysis of variance results for alpha for scores on work ethic scale in 2x2x2 experiment

Factors	Type III sum of squares	df	Mean Square	F-score	p-value	Partial eta-squared
Model	328.118	8	41.015	14081.526	.000	.996
Quality control items (Q)	.002	1	.002	.525	.469	.001
Confidential data (C)	.067	1	.067	22.864	.000	.051
Grouped items (G)	.106	1	.106	36.450	.000	.079
Q x C	.078	1	.078	26.708	.000	.059
G x Q	.047	1	.047	16.071	.000	.036
G x C	.016	1	.016	5.601	.018	.013
Q x C x G	.003	1	.003	1.029	.311	.002
Error	1.244	427	.003			
Total	329.362	435				

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 were 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* (Grissom & 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 the 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 in exchange for 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, reminders about the nature of coefficient alpha are warranted. 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

	Presence of quality control items		
Absence of quality control items	1.	2.	3.
1. Conscientiousness	--	-.360 (-.486, -.212)	.325 (.198, .443)
2. Entitlement	-.445 (-.557, -.327)	--	-.262 (-.387, -.134)
3. Work ethic	.337 (.208, .447)	-.202 (-.317, -.066)	--
	Items grouped together		
Items randomly scrambled	1.	2.	3.
1. Conscientiousness	--	-.529 (-.714, -.250)	.426 (.156, .625)
2. Entitlement	-.463 (-.652, -.241)	--	-.237 (-.484, .017)
3. Work ethic	.331 (.065, .536)	.004 (-.262, .268)	--
	Confidential data collection		
Anonymous data collection	1.	2.	3.
1. Conscientiousness	--	-.489 (-.692, -.228)	.332 (.081, .570)
2. Entitlement	-.463 (-.646, -.254)	--	-.171 (-.409, .093)
3. Work ethic	.331 (.062, .534)	.004 (-.266, .284)	--

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

REFERENCES

Akbulut, Y. (2015). Predictors of inconsistent responding in web surveys. *Internet Research*, *25*(1), 131–147. https://doi.org/10.1108/intr-01-2014-0017

Beach, D. A. (1989). Identifying the random responder. *The Journal of Psychology*, *123*(1), 101–113. https://doi.org/10.1080/00223980.1989.10542966

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, *17*(1), 10–17. https://doi.org/10.1111/j.1745-3992.1998.tb00616.x

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-experimental Designs for Research*. Houghton Mifflin Company.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd.). Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Costa, P. T. Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Psychological Assessment Resources, Inc.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596–612. https://doi.org/10.1177/0013164410366686

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/bf02310555

Desimone, J. A., Harms, P. D., Desimone, A. J., & Wood, D. (2018). The Differential Impacts of Two Forms of Insufficient Effort Responding. *Applied Psychology: An International Review*, *67*(2), 309–338. https://doi.org/10.1111/apps.12117

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Franke, G. H. (1997). "The Whole is More than the Sum of its Parts": The Effects of Grouping and Randomizing Items on the Reliability and Validity of Questionnaires. *European Journal of Psychological Assessment*, *13*(2), 67–74. https://doi.org/10.1027/1015-5759.13.2.67

Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7–28). Tillberg University Press.

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

Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, *41*(2), 219–231. https://doi.org/10.1007/bf02291840

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & Deshon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence constant dimensionality: the case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541. https://doi.org/10.1177/1094428115571894

Lim, L., Browder, D. M., & Sigafoos, J. (1998). The role of response effort and motion study in functionally equivalent task designs and alternatives. *Journal of Behavior and Education*, *81*, 81–102.

Lopez, M. N., & Charter, R. A. (2001). Random responding to the MMPI-2 F scale. *Psychological Reports*, *88*(2), 398–398. https://doi.org/10.2466/pr0.2001.88.2.398

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

McGrath, J. E. (1981). Dilemmatics: The study of research choices and dilemmas. *American Behavioral Scientist*, *25*(2), 179–210. https://doi.org/10.1177/000276428102500205

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Melnick, S. A. (1993). The effects of item grouping on the reliability and scale scores of an affective measure. *Educational and Psychological Measurement*, *53*(1), 211–216. https://doi.org/10.1177/0013164493053001023

Meyer, J. F., Faust, K. A., Faust, D., Baker, A. M., & Cook, N. E. (2013). Careless and Random Responding on Clinical and Research Measures in the Addictions: A Concerning Problem and Investigation of their Detection. *International Journal of Mental Health and Addiction*, *11*(3), 292–306. https://doi.org/10.1007/s11469-012-9410-5

Miller, B. K. (2009). Confirmatory factor analysis of the equity preference questionnaire. *Journal of Managerial Psychology*, *24*(4), 328–347. https://doi.org/10.1108/02683940910952714

Miller, M. J., Woehr, D. J., & Hudspeth, N. (2002). The meaning and measurement of work ethic: Construction and initial validation of a multidimensional inventory. *Journal of Vocational Behavior*, *60*(3), 451–489. https://doi.org/10.1006/jvbe.2001.1838

Morey, L. C., & Hopwood, C. J. (2004). Efficiency of a Strategy for Detecting Back Random Responding on the Personality Assessment Inventory. *Psychological Assessment*, *16*(2), 197–200. https://doi.org/10.1037/1040-3590.16.2.197

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, *45*(2), 239–250. https://doi.org/10.1002/1097-4679(198903)45:2

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). Guilford.

Pedhazur, E. J., & Schmelkin, L. P. (2013). *Measurement, design, and analysis: An integrated approach*. Psychology Press. https://doi.org/10.4324/9780203726389

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Podsakoff, P. M., & Organ, D. W. (1986). Self-reports in organizational research: Problems and prospects. *Journal of Management*, *12*(4), 531–544. https://doi.org/10.1177/014920638601200408

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized Conditional SEM: A Case for Conditional Reliability. *Applied Psychological Measurement*, *31*(3), 169–180. https://doi.org/10.1177/0146621606291569

Rush, M. C., Phillips, J. S., & Lord, R. G. (1981). Effects of a temporal delay in rating on leader behavior descriptions: A laboratory investigation. *Journal of Applied Psychology*, *66*(4), 442–450. https://doi.org/10.1037/0021-9010.66.4.442

Sauley, K. S., & Bedeian, A. G. (2000). Equity sensitivity: Construction of a measure and examination of its psychometric properties. *Journal of Management*, *26*(5), 885–910. https://doi.org/10.1177/014920630002600507

Schell, K. L., & Oswald, F. L. (2013). Personality and individual differences item grouping and item randomization. *Personality and Individual Differences*, *55*(3), 317–321. https://doi.org/10.1016/j.paid.2013.03.008

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measures*, *9*(4), 367–373. https://doi.org/10.1177/014662168500900405

Schriesheim, C. A., & DeNisi, A. S. (1980). Item Presentaton as an Influence on Questionnaire Validity: A Field Experiment. *Educational and Psychological Measurement*, *40*(1), 175–182. https://doi.org/10.1177/001316448004000130

Schriesheim, C. A., Kopelman, R. E., & Solomon, E. (1989). The Effect of Grouped versus Randomized Questionnaire Format on Scale Reliability and Validity: A Three-Study Investigation. *Educational and Psychological Measurement*, *49*(3), 487–508. https://doi.org/10.1177/001316448904900301

Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (2000). *The science of self-report*. Lawrence Erlbaum Associates.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837–847.

Toro-Zambrana, W., Lee, D. L., & Belfiore, P. J. (1999). The effects of response effort and choice on productivity. *Journal of Developmental Disabilities*, *11*(3), 1–7.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Ward, M. K., & Pond, S. B. I. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, *48*, 554–568. https://doi.org/10.1016/j.chb.2015.01.070

Woods, C. M. (2006). Careless Responding to Reverse-Worded Items: Implications for Confirmatory Factor Analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. https://doi.org/10.1007/s10862-005-9004-7

SUPPLEMENTARY MATERIALS

Peer Review Comments

Download: https://collabra.scholasticahq.com/article/17975-impact-of-survey-design-features-on-score-reliability/attachment/46862.docx?auth_token=RHwsoMFOKAJiC_LmiMlV