Does the Delivery Matter? Examining Randomization at the Item Level

Abstract

Scales that are psychometrically sound, meaning those that meet established standards regarding reliability and validity when measuring one or more constructs of interest, are customarily evaluated based on a set modality (i.e., computer or paper) and administration (fixed-item order). Deviating from an established administration profile could result in non-equivalent response patterns, indicating the possible evaluation of a dissimilar construct. Randomizing item administration may alter or eliminate these effects. Therefore, we examined the differences in scale relationships for randomized and nonrandomized computer delivery for two scales measuring meaning/purpose in life. These scales have questions about suicidality, depression, and life goals that may cause item reactivity (i.e. a changed response to a second item based on the answer to the first item). Results indicated that item randomization does not alter scale psychometrics for meaning in life scales, which implies that results are comparable even if researchers implement different delivery modalities.

Keywords: scales, randomization, item analysis

Does the Delivery Matter? Examining Randomization at the Item Level

The use of the Internet has been integrated into daily life as a means of accessing information, interacting with others, and tending to required tasks. The International Telecommunication Union reports that over half the world is online, and 70% of 15-24 year olds are on the internet (**Sanou2017**). Further, the Nielson Total Audience report from 2016 indicates that Americans spend nearly 11 hours a day in media consumption (**Media2016**). Researchers discovered that online data collection can be advantageous over laboratory and paper data collection, as it is often cheaper and more efficient (**Ilieva2001**; **Schuldt1994**; **Reips2012**). Internet questionnaires first appeared in the early 90s when HTML scripting code integrated form elements, and the first experiments appeared soon after (**Musch2000**; **Reips2002a**). The first experimental lab on the internet was the Web Experimental Psychology Lab formed by Reips (http://www.wexlab.eu), and the use of the Internet to collect data has since grown rapidly (**Reips2002a**). What started with email and HTML forms has since moved to whole communities of available participants including websites like Amazon's Mechanical Turk and Qualtrics' Participant Panels. Participants of all types and forms are easily accessible for somewhat little to no cost.

Our ability to collect data on the Internet has inevitably lead to the question of measurement invariance between in person and online data collection methods (**Meyerson2003**; **Buchanan2005**). Invariance implies that different forms, data collection procedures, or even target demographics produce comparable sets of responses, which is a desirable characteristic to ensure a minimal number of confounding variables (**Brown2006**). According to **Deutskens2006** mail surveys and online surveys produce nearly identical results regarding the accuracy of the data collected online versus by mail. Only minor differences arise between online surveys and mail in surveys when it comes to participant honesty and suggestions. For example, participants who responded to surveys online provided more suggestions, lengthier answers, and greater information about competitors in the field that they may prefer (**Deutskens2006**). The hypothesis as to why individuals may be more honest online than in person is that the individual may feel more anonymity and less social desirability effects due to the nature the online world, therefore less concerned about responding in a socially polite way (**Joinson1999**). A trend found by **Fang2012a** shows individuals are more likely to respond to surveys online with extreme scores, rather than mid-range responses on Likert scales due to the lessened social desirability factor. There may be slight cultural differences in responses online. For example, collectivistic cultures showed greater tendency toward mid-range responses on Likert scales via in-person and online due to placing greater value on how they are socially perceived; however, the trend is still the same as scores are more extreme online versus in person or by mail (**Fang2012**).

Although work by Dillman and his group (**Frick2001**; **Smyth2006**; **Dillman2008**), among others, has shown that many web surveys are plagued by problems of usability, display, coverage, sampling, non-response, or technology, other studies have found internet data to be reliable and almost preferable as it produces a varied demographic response compared to the traditional sample of introduction to psychology college students while also maintaining data equivalence (**Lewis2009**). However, equivalence in factor structure may be problematic, as **Buchanan2005** have shown that factor structure was not replicable in online and in person surveys. Other work has shown equivalence using a comparison of correlation matrices (**Meyerson2003**) or *t*-tests (**Schulenberg1999**; **Schulenberg2001**), and the literature is mixed on how different methodologies impact factor structure. **Weigold2013** recently examined both quantitative and research design questions (i.e. missing data) on Internet and paper-and-pencil administration which showed that the administrations were generally equivalent for quantitative structure but research design issues showed non-equivalence. Other potential limitations to online surveys include the accessibility of different populations to the Internet (**Frick2001**), selection bias (**Bethlehem2010**), response rates (**Cook2000**; **Hox1994**; **DeLeeuw1988**; **Cantrell2007**), attrition (**Cronk2002**), and distraction (**Tourangeau1999**). Many of these concerns have been alleviated in the years since online surveys were first developed, especially with the advent of panels and Mechanical Turk to reach a large, diverse population of participants (**Buhrmester2011**).

With the development of advanced online survey platforms such as Qualtrics and Survey Monkey, researchers have the potential to control potentially confounding research design issues through randomization, although other issues may still be present, such as participant misbehavior (**Nosek2002**). Randomization has been a hallmark of good research practice, as the order or presentation of stimuli can be a noise variable in a study with multiple measures (**Keppel2004**). Thus, researchers have often randomized scales by rotating the order of presentation in paper format or simply clicking the randomization button for web-based studies. This practice has counterbalanced out any order effects of going from one scale to the next (**Keppel2004**). However, while scale structure has remained constant, these items are still stimuli within a larger construct. Therefore, these construct-related items have the ability to influence the items that appear later on the survey, which we call item reactivity. For example, a question about being *prepared for death* or *thoughts about suicide* might change the responses to further questions, especially if previous questions did not alert participants to be prepare for that subject matter.

Scale development typically starts with an underlying latent variable that a researcher wishes to examine through measured items or questions (**DeVellis2016a**). Question design is a well-studied area that indicates that measurement is best achieved through questions that are direct, positively worded, and understandable to the subject (**Dillman2008**). **Olson2010** suggests researchers design a multitude of items in order to investigate and invite subject matter experts to examine these questions. Subject matter experts were found to be variable in their agreement, but excellent at identifying potentially problematic questions. After suggested edits from these experts, a large sample of participant data is collected. While item response theory is gaining traction, classical test theory has dominated this area through the use of exploratory and confirmatory factor analysis (**Worthington2006**). EFA elucidates several facets of how the measured items represent the latent trait through factor loadings and overall model fit (**Tabachnick2012**). Factor loadings represent the correlation between each item and the overall latent variable, where a researcher wishes to find items that are strongly related to the latent trait. Items that are not related to the latent trait, usually with factor loadings below .300 (**Preacher2003**) are discarded. Model fit is examined when simple structure has been achieved (i.e. appropriate factor loadings for each item), and these fit indices inform if the items and factor structure model fit the data well. Well-designed scales include items that are highly related to their latent trait and have excellent fit indices. Scale development additionally includes the examination of other measures of reliability (alpha) and construct validity (relation to other phenomena) but the focus of the scale shifts to subscale or total scores (**Buchanan2014**). Published scales are then distributed for use in the form that is presented in the publication, as item order is often emphasized through important notes about reverse scoring and creating subscale scores.

The question is no longer whether web-based surveys are reliable sources of data collection; the theory now is in need of a shift to whether or not item-randomization in survey data collection creates psychometric differences. These scale development procedures focus on items, and EFA/CFA statistically try to mimic variance-covariance structure by creating models of the data with the same variance-covariance matrix. If we imagine that stimuli in a classic experimental design can influence the outcome of a study because of their order, then certainly the stimuli on a scale (i.e., the items) can influence the pattern of responses for items. This area of study is relatively unexplored, as easily randomizing items has only recently become available for researchers.

Therefore, this study focuses on potential differences in results based on item randomization delivery methodology. The current project examined large samples on two logotherapy-related scales, as these scales include potentially reactive items, as well as both a dichotomous True/False and traditional Likert format for the same items. Large samples were desirable to converge on a stable, representative population; however, false positives (i.e., Type I errors) can occur by using large N. Recent developments in the literature focusing on null hypothesis testing make it especially important to present potential alternatives to p-values (**Valentine2017**). While a large set of researchers have argued that the literature is full of Type I errors (**Benjamin2017**), and thus, the α value should be shifted lower (i.e., $p < .005$ for statistical significance), an equally large set of researchers counter this argument as unfounded and weak (**Lakens2017**). We provide multiple sources of evidence (p-values, effect sizes, Bayes Factors, and tests of equivalence) to determine if differences found are not only statistically significant, but also practically significant. In our study, we expand to item randomization for online based surveys, examining the impact on item loadings to their latent variable, variance-covariance structure, item means, and total scores again providing evidence of difference/non-difference from multiple statistical sources. Finally, we examine these scenarios with a unique set of scales that have both dichotomous True/False and traditional Likert formats to explore how the answer response options might impact any differences found between randomized and nonrandomized methodologies.

Method

Participants

The sample population consisted of undergraduate students at a large Midwestern University, placing the approximate age of participants at around 18-22. Table 1 includes the demographic information about all datasets. Only two scales were used from each dataset, as described below. Participants were generally enrolled in an introductory psychology course that served as a general education requirement for the university. As part of the curriculum, the students were encouraged to participate in psychology research programs, resulting in their involvement in this study. These participants were given course credit for their participation.

Materials

Of the surveys included within each larger study, two questionnaires were utilized: the Purpose in Life Questionnaire (**Crumbaugh1964**) and the Life Purpose Questionnaire (**Hutzell1988**).

The Purpose in Life Questionnaire. The PIL is a 20-item questionnaire that assesses perceived meaning and life purpose. Items are structured in a 7-point Likert type response format; however, each item has different anchoring points that focus on item content. Total scores are created by summing the items, resulting in a range of 20 to 140 for the overall score. The reliability for the scale is generally high, ranging from .70 to .90 (**Schulenberg2004**; **Schulenberg2010**). Previous work on validity for the PIL showed viable one- and two-factor models, albeit question loadings varied across publications (**Schulenberg2010**), and these fluctuating results lead to the development of a 4-item PIL short form (**Schulenberg2011**).

Life Purpose Questionnaire. The LPQ was modeled after the full 20-item PIL questionnaire, also measuring perceived meaning and purpose in life. The items are structured in a true/false response format, in contrast to the Likert response format found on the PIL. Each question is matched to the PIL with the same item content, altering the question to create binary answer format. After reverse coding, zero on an item would indicate low meaning, while one on an item would indicate high meaning. A total score is created by summing questions, resulting in a range from 0 to 20. In both scales, higher scores indicated greater perceived meaning in life. Reliability for this scale is also correspondingly high, usually in the .80 range (**Melton2008**; **Schulenberg2004**).

These two scales were selected because they contained the same item content with differing response formats, which would allow for cross comparisons between results for each scale.

Procedure

The form of administration was of interest to this study, and therefore, two formats were included: computerized administration in nonrandom order and computerized administration with a randomized question order. Computerized questionnaires were available for participants to access electronically, and they were allowed to complete the experiment from anywhere with the Internet through Qualtrics. To ensure participants were properly informed, both an introduction and a debriefing were included within the online form. Participants were randomly assigned to complete a nonrandomized or randomized version of the survey. Nonrandomized questionnaires followed the original scale question order, consistent with paper delivery format. A different group of participants were given each question in a randomized order within each scale (i.e. all PIL and LPQ questions will still grouped together on one page). Scales were randomized across participants for both groups. Once collected, the results were then amalgamated into a database for statistical analysis.

Results

Hypothesis and Data-Analytic Plan

Computer forms were analyzed by randomized and nonrandomized groups to examine the impact of randomization on covariance structure, factor loadings, item means, and total scores. We expected to find that these forms may potentially vary across covariance structure and item means, which would indicate differences in reactivity to questions (i.e. item four always has item three as a precursor on a nonrandom form, while item four may have a different set of answers when prefaced with other questions). Factor loadings were assessed to determine if differences in randomization caused a change in focus, such that participant interpretation of the item changed the relationship to the latent variable. However, we did not predict if values would change, as latent trait measurement should be consistent. Last, we examined total scores; however, it was unclear if these values would change. A difference in item means may result in changes in total scores, but may also result in no change if some item means decrease, while others increase.

Each hypothesis was therefore tested using four dependent measures. First, we examined the variance-covariance matrix for each type of delivery and compared the matrices to each other by using root mean squared error (RMSE). RMSE estimates the difference between covariance matrices and is often used in structural equation modeling to determine if models have good fit to the data. A criterion of $< .06$ for good fit, .06-.08 for acceptable fit, and $> .10$ for bad fit was used (**Hu1999**). This analysis was used to determine if the change in delivery changed the structure of the item relationships to each other (i.e. if the correlation matrices are different). RMSE values were calculated using the *monomvn* package in *R* (**Gramacy2010**).

We then calculated an exploratory factor analysis on both scales using one-factor models to examine the loading of each item on its latent trait. The PIL factor structure is contested (**Strack2009**) with many suggestions as to latent structure for one- and two-factor models. The LPQ has seen less research on factor structure (**Schulenberg2004**). This paper focused on loadings on one global latent trait to determine if the manipulation of delivery impacted factor loadings. We used a one-factor model and included all questions to focus on the loadings, rather than the factor structure. The analysis was performed using the *psych* package in *R* with maximum likelihood estimation and an oblique (oblimin) rotation. The LPQ factor analysis used tetrachoric correlation structure to control for the dichotomous format of the scale, rather than traditional Pearson correlation structure. The loadings were then compared using a matched dependent *t*-test (i.e. item one to item one, item two to item two) to examine differences between nonrandomized and randomized computer samples.

Next, item averages were calculated across all participants for each item. These 20 items were then compared in a matched dependent *t*-test to determine if delivery changed the mean of the item on the PIL or LPQ. While covariance structure elucidates the varying relations between items, we may still find that item averages are pushed one direction or another by a change in delivery and still maintain the same correlation between items. If this test was significant, we examined the individual items across participants for large effect sizes, as the large sample sizes in this study would create significant t-test follow ups.

Last, the total scores for each participant were compared across delivery type using an independent t-test. Item analyses allow a focus on specific items that may show changes, while total scores allow us to investigate if changes in delivery alter the overall score that is used in other analyses or possible clinical implications. For analyses involving t-tests, we provide multiple measures of evidentiary value so that researchers can weigh the effects of randomization on their own criterion. Recent research on α criteria has shown wide disagreement on the usefulness of p-values and set cut-off scores (**Benjamin2017**; **Lakens2017**). Therefore, we sought to provide traditional null hypothesis testing results (t-tests, p-values) and supplement these values with effect sizes (**Cumming2014**; **Buchanan2017**; **Smithson2001**), Bayes Factors (**Kass1995**; **Morey2015b**), and one-sided tests of equivalence (**Cribbie2004**; **Lakens2017a**; **Schuirmann1987**; **Rogers1993**). We used the average standard deviation of each group as the denominator for d calculation as follows:

$$d_{av} = \frac{(M_1 - M_2)}{\frac{SD_1 + SD_2}{2}}$$

This effect size is less biased than the traditional d_z formula, wherein mean differences are divided by the standard deviation of the difference scores (**Lakens2013**). The difference scores standard deviation is often much smaller than the average of the standard deviations of each level, which can create an upwardly biased effect size (**Cumming2014**). This bias can lead researchers to interpret larger effects for a psychological phenomenon than actually exist.

Bayes Factors are calculated in opposition to a normal frequentist (NHST) approach, as a ratio of the likelihood of two models. Traditional NHST focuses on the likelihood of the data, given the null hypothesis is true, and Bayesian analysis instead posits the likelihood of a hypothesis given the data. Prior distributions are our estimation of the likelihood of our hypothesis before the data was collected, which is combined with the data collected to form a posterior belief of our hypothesis. We chose to use Bayes Factors as a middle ground to the Bayesian analysis continuum, that uses mildly uninformative priors and allows for the data to strongly impact the posterior distribution. The choice of prior distribution can heavily influence the posterior belief, in that uninformative priors allow the data to comprise the posterior distribution. However, most researchers have a background understanding of their field, thus, making completely uninformative priors a tenuous assumption. Because of the dearth of literature in this field, there is not enough previous information to create a strong prior distribution, which would suppress the effect of the data on posterior belief. The *BayesFactor* package (**Morey2015b**) uses recommended default priors that cover a wide range of data (**Morey2015b**; **Rouder2009**) of a Jeffreys prior with a fixed rscale (0.5) and random rscale (1.0). The alternative model is generally considered a model wherein means between groups or items differ, and this model is compared to a null model of no mean differences. The resulting ratio is therefore the odds of the alternative model to the null, where BF values less than one indicate evidence for the null, values at one indicate even evidence for the null and alternative, and values larger than one indicate evidence for the alternative model. While some researchers have posed labels for BF values (**Kass1995**), we present these values as a continuum to allow researchers to make their own decisions (**Morey2015b**; **Morey2015c**).

NHST has also been criticized for an inability to test the null hypothesis, and thus, show evidence of the absence of an effect. Non-significant p-values are often misinterpreted as evidence for the null hypothesis (**Lakens2017a**). However, we can use the traditional frequentist approach to determine if an effect is within a set of equivalence bounds. We used the two one-sided tests approach to specify a range of raw-score equivalence that would be considered supportive of the null hypothesis (i.e. no worthwhile effects or differences). TOST are then used to determine if the values found are outside of the equivalence range.

Significant TOST values indicate that the effects are *within* the range of equivalence. We used the *TOSTER* package (**Lakens2017a**) to calculate these values, and graphics created from this package can be found online at https://osf.io/gvx7s/.

The equivalence ranges are often tested by computing an expected effect size of negligible range; however, the TOST for dependent t uses d_z, which can overestimate the effect size of a phenomena (**Cumming2014**; **Lakens2013**). Therefore, we calculated TOST tests on raw score differences to alleviate the overestimation issues. For EFA, we used a change score of .10 in the loadings, as Comrey and Lee (1992) suggested loading estimation ranges, such as .32 (poor) to .45 (fair) to .55 (good), and the differences in these ranges are approximately .10 (**Tabachnick2012**). Additionally, this score would amount to a small correlation change using traditional guidelines for interpretation of r (**Cohen1992a**). For item and total score differences, we chose a 5% change in magnitude as the raw score cut off as a modest raw score change. To calculate that change for total scores, we used the following formula:

$$(Max * N_{Questions} - Min * N_{Questions}) * Change$$

Minimum and maximum values indicate the lower and upper end of the answer choices (i.e. 1 and 7), and change represented the proportion magnitude change expected. Therefore, for total PIL scores, we proposed a change in 6 points to be significant, while LPQ scores would change 1 point to be a significant change. For item analyses, we divided the total score change by the number of items to determine how much each item should change to impact the total score a significant amount (PIL = 0.30, LPQ = .05).

Data Screening

Each dataset was analyzed separately by splitting on scale and randomization, and first, all data were screened for accuracy and missing data. Participants with more than 5% missing data (i.e. 2 or more items) were excluded. Data were imputed using the *mice* package in *R* for participants with less than 5% of missing data (**VanBuuren2011**). Next, each dataset was examined for multivariate outliers using Mahalanobis distance (**Tabachnick2012**). Each dataset was then screened for multivariate assumptions of additivity, linearity, normality, homogeneity, and homoscedasticity. While some data skew was present, large sample sizes allowed for the assumption of normality of the sampling distribution. Information about the number of excluded data points in each step is presented in Table 1.

PIL Analyses

Covariance Matrices. Covariance structure was considered different (**Hu1999**) for the randomized and not randomized forms of item order, $RMSE = .15$. Standardized residuals were calculated by dividing the difference in covariance tables by the variance of the differences (**Hausman1978**). While *RMSE* indicated partial misfit between the covariance relationships, only 3 values were significantly different using Z of 1.96 as a criterion: the variances of PIL 7 and 14. PIL 7 in a randomized form had less variance ($SD^2 = 1.46$) than the nonrandomized form ($SD^2 = 1.89$). Likewise, PIL 14 randomized had a smaller variance ($SD^2 = 1.90$) than the nonrandomized form ($SD^2 = 2.40$). Questions about retirement and freedom to make choices decreased in variance when they were randomly presented.

Factor Loadings. Table 2 includes the factor loadings from the one-factor EFA analysis. These loadings were compared using a dependent t-test matched on item, and they were not significantly different, $M_d = 0.00$, 95% CI $[-0.02, 0.03]$, $t(19) = 0.26$, $p = .801$. The effect size for this test was correspondingly negligible, d_{av} = -0.02 95% CI [-0.45, 0.42]. The TOST test was significant for both the lower, $t(19) = 0.19$, $p < .001$ and the upper bound, $t(19) =$ -0.70, $p < .001$. This result indicated that the change score was within the confidence band of expected negligible changes. Lastly, the BF for this test was 0.24 $\pm$0.02%, which indicated support for the null model.

Item Means. Table 2 includes the means and standard deviation of each item from the PIL scale. The item means were compared using a dependent t-test matched on item. Item means were significantly different $M_d = -0.07$, 95% CI $[-0.13, -0.02]$, $t(19) = -2.88$, $p = .010$. The effect size for this difference was small, d_{av} = -0.16 95% CI [-0.60, 0.29]. Even though the t-test was significant, the TOST test indicated that the difference was within the range of a 5% percent change in item means (0.30). The TOST test for lower bound, $t(19) =$ -1.54, $p < .001$ and the upper bound, $t(19)$ = -4.22, $p < .001$, suggested that the significant t-test may be not be interpreted as a meaningful change on the item means. The BF value for this test indicated 6.86 <0.01%, which is often considered weak evidence for the alternative model. Here, we find mixed results, indicating that randomization may change item means for the PIL.

Total Scores. Total scores were created by summing the items for each participant across all twenty PIL questions. The mean total score for nonrandomized testing was M = 103.00 (SD = 18.31), while the mean for randomizing testing was M = 104.46 (SD = 17.83). This difference was examined with an independent t-test and was not significant, $t(1,897) = -1.75$, $p = .081$. The effect size for this difference was negligible, d_{av} = -0.08 95% CI [-0.17, 0.29]. We tested if scores were changed by 5% (6.00 points), and the TOST test indicated that the lower, $t(1897)$ = 5.44, $p < .001$ and the upper bound, $t(1897)$ = -8.94, p < .001 were within this area of null change. The BF results also supported the null model, 0.25 <0.01%.

LPQ Analyses

Covariance Matrices. Covariance structure for the LPQ was found to be the same across both randomized and nonrandomized testing, $RMSE$ = .02. Standardized residuals indicated that the covariance between items 9 and 11 were significantly different, while item 13 included significantly different variances. The correlation between items 9 (empty life) and 11 (wondering about being alive) for randomized versions was r = .33 while the correlation for nonrandomized versions was $r = .51$. The variance for item 13 (responsibility) in a randomized version ($SD^2 = .03$) was smaller than the variance in the nonrandomized version ($SD^2 = .08$).

Factor Loadings. Table 3 includes the factor loadings from the one-factor EFA analysis using tetrachoric correlations. The loadings from randomized and nonrandomized versions were compared using a dependent t-test matched on item, which indicated they were not significantly different, $M_d = 0.01$, 95% CI $[-0.01, 0.04]$, $t(19) = 0.99$, $p = .336$. The difference found for this test was negligible, d_{av} = -0.07 95% CI [-0.50, 0.37]. The TOST test examined if any change was within .10 change, as described earlier. The lower, $t(19)$ = -0.54, $p < .001$ and the upper bound, $t(19)$ = -1.43, $p < .001$ were both significant, indicating that the change was within the expected change. Further, in support of the null model, the BF was 0.34 $\pm$0.02%.

Item Means. Means and standard deviations of each item are presented in Table 3. We again matched items and tested if there was a significant change using a dependent t-test. The test was not significant, $M_d = 0.00$, 95% CI $[-0.02, 0.02]$, $t(19) = 0.26$, $p = .801$, and the corresponding effect size reflects how little these means changed, d_{av} = 0.01 95% CI [-0.42, 0.45]. Using a 5% change criterion, items were tested to determine if they changed less than (0.05). The TOST test indicated both lower, $t(19) = 0.48$, $p < .001$ and the upper bound, $t(19) = 0.03$, $p < .001$, were within the null range. The BF also supported the null model, 0.24 $\pm$0.02%.

Total Scores. LPQ total scores were created by summing the items for each participant. The mean total score for randomized testing was $M = 14.14$ ($SD = 4.01$), and the mean for nonrandomized testing was $M = 14.19$ ($SD = 4.22$). An independent t-test indicated that the testing did not change total score, $t(1, 632) = 0.23$, $p = .819$. The effect size for this difference was negligible, d_{av} = 0.01 95% CI [-0.09, 0.45]. The TOST test indicated that the scores were withing a 5% (1.00 points) change, lower: $t(1627) = 5.13$, $p <$ .001 and upper: $t(1627)$ = -4.68, $p < .001$. The BF results were in support of the null model, 0.06 ±0.04%.

Discussion

As technology has advanced, initial research questioned the validity of online assessments versus paper assessments. With further investigation, several researchers discovered measurement invariance with regard to computer surveys compared with paper surveys (**Deutskens2006**; **Lewis2009**). However, with the addition of technology, **Fang2012a** suggested that individuals respond with more extreme scores in online surveys than in-person surveys due to the social-desirability effect. Research on scale invariance is mixed in results for paper and computer, and our work is a first-step on examining survey equivalence on an individual item-level for different forms of computer delivery.

The findings from the current study imply that item randomization is a viable option for controlling any potential reactivity between questions. First, as we analyzed the PIL, the covariance matrices were non-equivalent; the randomized data show decreased variance for several items compared to the nonrandomized data. Since variance provides a measure of how the data vary around the mean, decreased variance typically results in decreased measurement error; thus, randomization has the potential to decrease measurement error in data collection. The findings also support the null hypothesis in regards to factor loading differences because the item relationship to a latent variable should not change with randomization. The item means comparison resulted in significant differences between item randomization and nonrandomization using p-value criterion and Bayes Factor analyses. However, the effect size was small, meaning the differences were not as meaningful as the p-values and BF analyses posit, in addition to considering the evidentiary values of the two one-sided tests, which supported the null range of expected values. Finally, the total scores showed equivalence between randomization and nonrandomization which suggested that total scales were not considerably impacted with or without randomization of items.

Analyses for the LPQ yielded somewhat similar results to those of the PIL. Pertaining to covariance structures, the randomized and nonrandomized scales resulted in equivalence, with a recapitulation of the PIL analysis in which variance was decreased in the randomized sample for at least one item. A slight correlational difference was detected for items 9 and 11 in which the nonrandomized scale shows a large association between the items, while the randomized scale shows a moderate association between the items. However, the presence of the association remained present on both randomized and nonrandomized scales. Further analyses of the factor loadings, item means, and total scores resulted in equivalence between forms. Therefore, the null hypothesis was supported. Evidentiary equivalence for item means and total scores suggested that randomization of items was not disadvantaging the overall scoring structure of the scale and provides further support for randomization as a means of methodological control. The match between results for two types of answer methodologies (i.e. Likert and True/False) implied that randomization can be applied across a variety of scale types with similar effects.

Since the PIL and LPQ analyses predominately illustrated support for null effects of randomization, item randomization of scales is of practical use when there are potential concerns about item order. Randomization has been largely viewed as virtuous research practice in terms of sample selection and order of stimuli presentation for years; now, we must decide if item reactivity earns the same amount of caution that has been granted to existing research procedures. Since we found equivalence in terms of overall scoring of the PIL and LPQ, we advise that randomization should and can be used as a control mechanism, in addition to the ease of comparison between the scales if one researcher decided to randomize and one did not. Moreover, these results would imply that if an individual's total score on the PIL or LPQ is significantly different on randomized versus nonrandomized administrations, it is likely due to factors unrelated to delivery. Future research should investigate if this result is WEIRD (Western, Educated, Industrialized, Rich, and Democratic), as this study focused on college-age students in the Midwest (**Henrich2010**). As **Fang2012**'s research indicates different effects for collectivistic cultures, other cultures may show different results based on randomization. Additionally, one should consider the effects of potential computer illiteracy on online surveys (**Charters2004**).

A second benefit to using the procedures outlined in this paper to examine for differences in methodology is the simple implementation of the analyses. While our analyses were performed in *R*, nearly all of these analyses can be performed in free point and click software, such as *jamovi* and *JASP*. Multigroup confirmatory factory analyses can additionally be used to analyze a very similar set of questions (**Brown2006**); however, multigroup analyses require a specialized skill and knowledge set. Bayes Factor and TOST analyses are included in these free programs and are easy to implement. In this paper, we have provided examples of how to test the null hypothesis, as well as ways to include multiple forms of evidentiary value to critically judge an analysis on facets other than p-values (**Valentine2017**).

References

Table 1

Demographic and Data Screening Information

Group	Female	White	Age (SD)	Original N	Missing N	Outlier N
PIL Random	61.6	81.1	19.50 (2.93)	1462	333	58
PIL Not Random	54.1	78.6	19.68 (3.58)	915	51	36
LPQ Random	-	-	-	1462	555	23
LPQ Not Random	-	-	-	915	150	15

Note. Participants took both the PIL and LPQ scale, therefore, random and not random demographics are the same. Not every participant was given the LPQ, resulting in missing data for those subjects. Several PIL participants were removed because they were missing an item on their scale.

Table 2

Item Statistics for the PIL Scale

Item	FL-R	FL-NR	M-R	SD-R	M-NR	SD-NR
1	.671	.638	4.825	1.278	4.806	1.278
2	.678	.573	4.928	1.438	4.600	1.452
3	.685	.671	5.811	1.126	5.732	1.101
4	.840	.846	5.675	1.302	5.655	1.285
5	.637	.574	4.669	1.495	4.409	1.497
6	.675	.684	5.421	1.314	5.338	1.400
7	.422	.439	6.174	1.207	6.081	1.373
8	.628	.598	5.014	1.092	5.010	1.138
9	.823	.796	5.355	1.177	5.327	1.198
10	.720	.765	5.209	1.494	5.155	1.544
11	.776	.796	5.227	1.621	5.163	1.623
12	.604	.648	4.494	1.568	4.522	1.601
13	.428	.402	5.745	1.243	5.737	1.216
14	.450	.421	5.427	1.380	5.240	1.548
15	.081	.221	4.375	1.940	4.147	1.885
16	.553	.554	5.088	1.991	5.267	1.862
17	.722	.735	5.418	1.396	5.395	1.403
18	.481	.501	5.384	1.474	5.302	1.593
19	.680	.720	4.878	1.416	4.905	1.454
20	.781	.811	5.343	1.313	5.210	1.289

Note. FL = Factor Loadings, M = Mean, SD = Standard Deviation, R = Random, NR = Not Random

Table 3

Item Statistics for the LPQ Scale

Item	FL-R	FL-NR	M-R	SD-R	M-NR	SD-NR
1	.676	.681	.567	.496	.613	.487
2	.901	.869	.755	.431	.761	.427
3	.503	.397	.864	.343	.843	.364
4	.725	.686	.907	.290	.868	.339
5	.689	.685	.419	.494	.509	.500
6	.511	.560	.637	.481	.581	.494
7	.189	.287	.774	.419	.811	.392
8	.557	.473	.483	.500	.467	.499
9	.856	.909	.812	.391	.781	.414
10	.594	.620	.636	.481	.647	.478
11	.639	.756	.727	.446	.760	.427
12	.683	.756	.786	.410	.751	.433
13	.314	.401	.965	.184	.909	.287
14	.484	.481	.761	.427	.769	.422
15	.050	.101	.322	.468	.395	.489
16	.697	.705	.862	.345	.872	.334
17	.517	.505	.848	.359	.813	.390
18	.559	.513	.829	.377	.828	.378
19	.675	.713	.464	.499	.497	.500
20	.636	.616	.723	.448	.712	.453

Note. FL = Factor Loadings, M = Mean, SD = Standard Deviation, R = Random, NR = Not Random