Does the Delivery Matter? Examining Randomization at the Item Level

Erin M. Buchanan[1], Riley E. Foreman[1], Becca N. Johnson[1], Jeffrey M. Pavlacic[2], Rachel L. Swadley[1], & Stefan E. Schulenberg[2]

[1] Missouri State University

[2] University of Mississippi

Author Note

Erin M. Buchanan is an Associate Professor of Quantitative Psychology at Missouri State University. Riley E. Foreman received his undergraduate degree in Psychology and Cell and Molecular Biology at Missouri State University and is currently at Kansas City University of Medicine and Biosciences. Becca N. Johnson is a masters degree candidate at Missouri State University. Jeffrey M. Pavlacic is a doctoral candidate at The University of Mississippi. Rachel N. Swadley completed her master's degree in Psychology at Missouri State University. Stefan E. Schulenberg is a Professor of Clinical Psychology at The University of Mississippi and Director of the Clinical Disaster Research Center. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Correspondence concerning this article should be addressed to Erin M. Buchanan, 901 S. National Ave. E-mail: erinbuchanan@missouristate.edu

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 (Sanou, 2017). Further, the Nielson Total Audience report from 2016 indicates that Americans spend nearly 11 hours a day in media consumption (Media, 2016). Researchers discovered that online data collection can be advantageous over laboratory and paper data collection, as it is often cheaper and more efficient (Ilieva, Baron, & Healy, 2002; Reips, 2012; Schuldt & Totten, 1994). Internet questionnaires first appeared in the early 90s when HTML scripting code integrated form elements, and the first experiments appeared soon after (Musch & Reips, 2000; Reips, 2002). 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 (Reips, 2002). 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 equivalence between in person and online data collection methods (Buchanan et al., 2005; Meyerson & Tryon, 2003). We will use the term equivalence as a global term for measurement of the same underlying construct between groups, forms, or testing procedures given no other manipulations. A related concept is measurement invariance, which focuses on the statistical and psychometric structure of measurement (Brown, 2006; Meredith, 1993). Multigroup confirmatory factor analysis (MGCFA) and multiple-indicators-multiple causes (MIMIC) structural models are often used to explore invariance in groups (Brown, 2006; Steenkamp & Baumgartner, 1998). The general approach through MGCFA explores if the latent structure of the proposed model is similar across groups (equal form or configural invariance), followed by more stringent tests indicating equal factor loadings (metric invariance), equal item intercepts (scalar invariance), and potentially, equal error variances (strict invariance). These steps can be used to determine where and how groups differ when providing responses to questionnaires and to propose changes to interpretations of test scores (for an example, see Trent et al., 2013). Measurement invariance implies equivalence between examined groups, while overall equivalence studies may not imply the psychometric concept of invariance.

Research has primarily focused on simple equivalence, with more uptick in research that specifically focuses on measurement invariance with the advent of programs that make such procedures easier. When focusing on equivalence, Deutskens, de Ruyter, and Wetzels (2006) found that 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 (Deutskens et al., 2006). 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 of the online world, therefore less concerned about responding in a socially polite way (Joinson, 1999). A trend found by Fang, Wen, and Pavur (2012a) shows individuals are more likely to respond to surveys online with extreme scores, rather than mid-range responses on 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 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 (Fang, Wen, & Prybutok, 2012b).

Although work by Dillman and his group (Dillman, Smyth, & Christian, 2008; Frick, Bächtiger, & Reips, 2001; Smyth, 2006), 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 equivalence (Lewis, Watson, & White, 2009). However, equivalence in factor structure may be problematic, as Buchanan et al. (2005) 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 (Meyerson & Tryon, 2003) or *t*-tests (Schulenberg & Yutrzenka, 1999, 2001), and the literature is mixed on how different methodologies impact factor structure. Weigold, Weigold, and Russell (2013) 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 (Frick et al., 2001), selection bias (Bethlehem, 2010), response rates (Cantrell & Lupinacci, 2007; Cook, Heath, & Thompson, 2000; De Leeuw & Hox, 1988; Hox & De Leeuw, 1994), attrition (Cronk & West, 2002), and distraction (Tourangeau, Rips, & Rasinski, 1999). 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 (Buhrmester, Kwang, & Gosling, 2011).

With the development of advanced online survey platforms such as Qualtrics and Survey Monkey, researchers have the potential to control for confounding research design issues through randomization, although other issues may still be present, such as participant misbehavior (Nosek, Banaji, & Greenwald, 2002). 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 (Keppel & Wickens, 2004). 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 (Keppel & Wickens, 2004). 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 prepared for that subject matter.

Scale development typically starts with an underlying latent variable that a researcher wishes to examine through measured items or questions (DeVellis, 2016). 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 (Dillman et al., 2008). Olson (2010) 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 (EFA, CFA; Worthington & Whittaker, 2006). EFA elucidates several facets of how the measured items represent the latent trait through factor loadings (Tabachnick & Fidell, 2012). Factor structure represents the correlation between item scores and factors, where a researcher wishes to find items that are strongly related to latent traits. Items that are not related to the latent trait, usually with factor loadings below .300 (Preacher & MacCallum, 2003) are discarded. Well-designed scales include items that are highly related to their latent trait. Scale development additionally includes the examination of other measures of reliability and validity but the focus of the scale shifts to subscale or total scores (Buchanan, Valentine, & Schulenberg, 2014). 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. Measuring an attitude or phenomena invokes a reaction in the participant (Knowles et al., 1992). Often, this reaction or reactivity is treated as error in measurement, rather than a variable to be considered in the experiment (Webb, Campbell, Schwartz, & Sechrest, 1966). Potentially, reaction to items on a survey could integrate self-presentation or social desirability (Webb et al., 1966) but cognitive factors also contribute to the participant response. Rogers (1974) and Tourangeau and Rasinski (1988) suggested a four part integration process that occurs when responses are formulated to questions. First, the participant must interpret the item. The interpretation process usually allows for one construal, and other interpretations may be ignored (Lord, Lepper, & Preston, 1984). Based on this process, information about the item must be pulled from memory. The availability heuristic will bias information found for the next stage, the judgment process, especially given the mood of the participant (MacLeod & Campbell, 1992; Tversky & Kahneman, 1973). These memories and information, by being recalled as part of answering an item, are often strengthened for future judgments or recall (Bargh & Pratto, 1986; Posner, 1978).

The judgment process has important consequences for the answers provided on a questionnaire. Judgments are often polarized because of the cognitive processes used to provide that answer (Tesser, 1978). The participant may become more committed to the answer provided (Feldman & Lynch, 1988), and future judgments are “anchored” against this initial judgment (Higgins & Lurie, 1983; Strack, Schwarz, & Gschneidinger, 1985). Finally, future memory searches will be confirmatory for the judgment decision (Petty & Cacioppo, 1986). The response selection is the final stage of the Rogers (1974) and Tourangeau and Rasinski (1988) models. This model provides an excellent framework through which to view the consequences of merely being asked a question. In this study, the focus is on the final stage of response selection, as it is the recordable output of these cognitive processes. Knowles et al. (1992) discuss that the item order may create a context effect for each subsequent question, wherein participants are likely to confuse the content of an item with the context of the previous questions. Their meaning-change hypothesis posits that each following item will be influenced by the previous set of items and does have important consequences for the factor loadings and reliability of the scale. Indeed, Salancik and Brand (1992) indicate that item order creates a specific context that integrates with background knowledge during the answering process, which can create ambiguity in measurement of the interested phenomenon. Panter, Tanaka, and Wellens (1992) discuss these effects from classic studies of item ordering, wherein agreement to a specific item first reduces agreement to a more general item second (Strack & Martin, 1987).

Given this previous research on item orderings, this study focuses on potential differences in results based on item randomization delivery methodology. This work is especially timely given the relative ease with which randomization can be induced with survey software. The current project examined large samples on two logotherapy-related scales, as these scales include potentially reactive items (e.g., death and suicide items embedded in positive psychology questions), as well as both a dichotomous True/False and traditional 1-7 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 (Valentine, Buchanan, Scofield, & Beauchamp, 2017). While a large set of researchers have argued that the literature is full of Type I errors (Benjamin et al., 2018), 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 (Lakens et al., 2018). 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 factor loadings, correlation 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 1-7 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 (PIL; Crumbaugh & Maholick, 1964) and the Life Purpose Questionnaire (LPQ; Hutzell, 1988).

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 type response format; however, each item has different anchoring points that focus on item content. No items are reverse scored, although, items are presented such that the 7 point end would be equally presented on the left and right when answering. Therefore, these items would need to be reverse coded if computer software automatically codes each item from 1 to 7 in a left to right format. Total scores are created by summing the items, resulting in a range of 20 to 140 for the overall score. The reliability reported for the scale has previously ranged from .70 to .90 (Schulenberg, 2004; Schulenberg & Melton, 2010). Previous work on validity for the PIL showed viable one- and two-factor models, albeit factor loadings varied across publications (see Schulenberg & Melton, 2010 for a summary), and these fluctuating results lead to the development of a 4-item PIL short form (Schulenberg, Schnetzer, & Buchanan, 2011).

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 1-7 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, scoring a zero on an item would indicate low meaning, while scoring a one on an item would indicate high meaning. A total score is created by summing item scores, resulting in a range from 0 to 20. In both scales, higher scores indicated greater perceived meaning in life. Reliability reported for this scale is usually in the .80 range (Melton & Schulenberg, 2008; Schulenberg, 2004).

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). The order of administration of the two scales was randomized across participants for both groups. Once collected, the results were then amalgamated into a database for statistical analysis.

Results

Hypotheses and Data-Analytic Plan

Computer forms were analyzed by randomized and nonrandomized groups to examine the impact of randomization on equivalence through correlation matrices, factor loadings, item means, and total scores. We expected to find that these forms may potentially vary across correlation structure and item means, which would indicate differences in reactivity and item context 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; Knowles et al., 1992). Factor loadings were assessed to determine if differences in randomization caused a change in loadings (Buchanan et al., 2005). However, we did not predict if these values would be different, as previous research indicates that participants may have a change in context with a different item order, but this change may not impact the items relationship with the factor. 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 correlation matrix for each type of delivery and compared the matrices to each other by using the *cortest.mat* function in the *psych* package (Revelle, 2017). This test provides a χ^2 value that represents the difference between a pair of correlation matrices. If this value was significant, we followed up by exploring the differences between correlations individually using Fisher's r to z transformation. Each pair of correlations (i.e., random r_{12} versus nonrandom r_{12}) was treated as an independent correlation and the difference between them was calculated by:

$$Z_{difference} = \frac{(Z_1 - Z_2)}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

Critical $Z_{difference}$ was considered +/- 1.96 for this analysis, and all values are provided online on at https://osf.io/gvx7s/. This manuscript was written in *R* markdown with the *papaja* package (Aust & Barth, 2017), and this document, the data, and all scripts used to calculate our statistics are available on the OSF page.

We then conducted 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 (Strack & Schulenberg, 2009) with many suggestions as to latent structure for one- and two-factor models. The LPQ has seen less research on factor structure (Schulenberg, 2004). 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. 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 correlation 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 (Benjamin et al., 2018; Lakens et al., 2018). Therefore, we sought to provide traditional null hypothesis testing results (*t*-tests, *p*-values) and supplement these values with effect sizes (*d* and non-central confidence intervals, Buchanan, Valentine, & Scofield, 2017; Cumming, 2014; Smithson, 2001), Bayes Factors (Kass & Raftery, 1995; Morey & Rouder, 2015), and two one-sided tests of equivalence (TOST, Cribbie, Gruman, & Arpin-Cribbie, 2004; Lakens, 2017; Rogers, Howard, & Vessey, 1993; Schuirmann, 1987).

For dependent *t*-tests, we used the average standard deviation of each group as the denominator for *d* calculation as follows (Cumming, 2012):

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

This effect size for repeated measures was used instead of the traditional d_z formula, wherein mean differences are divided by the standard deviation of the difference scores (Lakens, 2013). 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 (Cumming, 2014). This bias can lead researchers to interpret larger effects for a psychological phenomenon than actually exist. Lakens (2013) recommends using d_{av} over d_z because d_z can overestimate the effect size (see also, Dunlap, Cortina, Vaslow, & Burke, 1996) and d_{av} can be more comparable to between subjects designs d values. For independent t-tests, we used the d_s formula (Cohen, 1988):

$$d_s = \frac{(M_1 - M_2)}{\sqrt{\frac{(N_1-1)SD_1+(N_2-1)SD_2}{N_1+N_2-2}}}$$

The normal frequentist approach (NHST) focuses largely on significance derived from p-values while Bayesian approaches allow for the calculation of Bayes Factors that provide estimates of the support for one model as compared to another (Dienes, 2014; Wagenmakers, 2007). NHST methods traditionally involve two competing hypotheses: a null or nil hypothesis of no change between groups (Cohen, 1994) and an alternative or research hypothesis of change between groups, as a mish-mash of Fisherian and Neyman-Pearson methods. However, one limitation to this approach is the inability to support the null hypothesis (Gallistel, 2009). Within a Bayesian framework, one focuses on the uncertainty or probability of phenomena, including the likelihood of no differences between groups (Lee & Wagenmakers, 2014). Again, we can create two models: one of the null where both groups arise from the distribution with given parameters and one of the alternative where each group arises from different distributions with their own unique parameters. For both these models, before seeing the data, the researcher decides what they believe the distributions of these parameters look like before creating prior distributions. When data is collected, it is used to inform and update these prior distributions creating posterior distributions. Because the Bayesian framework focuses on updating previous beliefs with the data collected to form new beliefs, any number of hypotheses may be tested (for a humorous example, see Wagenmakers, Morey, & Lee, 2016). A Bayesian version of significance testing may be calculated by using model comparison through Bayes Factors (Etz & Wagenmakers, 2017; Kass & Raftery, 1995; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Bayes Factors are calculated as a ratio of the marginal likelihood of the two models. Bayes Factors provide a numeric value for how likely one model is over another model, much like likelihood or odds ratios.

Here, Bayes Factors (BF) are calculated as the marginal likelihood of the observed data under the alternative hypothesis divided by the marginal likelihood of the data with the null hypothesis. The resulting ratio can therefore give evidence to the support of one model as compared to another, where BF values less than one indicate support for the null model, values near one indicate both models are equally supported, and values larger than one indicate support for the alternative model. While some researchers have proposed conventions for BF values to discuss the strength of the evidence (Kass & Raftery, 1995), we will present these values as a continuum to allow researchers to make their own decisions (Morey, 2015; Morey & Rouder, 2015). Using this Bayesian approach, we are then able to show support for or against the null model, in contrast to NHST where we can only show support against the null (Gallistel, 2009).

Specifically, we used the *BayesFactor* package (Morey & Rouder, 2015) with the recommended default priors that cover a wide range of data (Ly, Verhagen, & Wagenmakers, 2016; Morey & Rouder, 2015; Rouder et al., 2009) of a Jeffreys prior with a fixed rscale (0.5) and random rscale (1.0). 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. Therefore, we used the default options in *BayesFactor* to model this belief.

Using Bayes Factors, we may be able to show evidence of the absence of an effect. Often, non-significant *p*-values from a NHST analysis are misinterpreted as evidence for the null hypothesis (Lakens, 2017). 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 (TOST) 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 (Lakens, 2017) to calculate these values, and graphics created from this package can be found online on our OSF page.

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 (Cumming, 2014; Lakens, 2013). Therefore, we calculated TOST analyses 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 (as cited in Tabachnick & Fidell, 2012, p. 654). Additionally, this score would amount to a small correlation change using traditional guidelines for interpretation of r (Cohen, 1992). 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 need to change 1 point to be significant. 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).

As discussed in the introduction, another approach to measuring equivalence would be through a MGCFA framework, analyzing measurement invariance. Those analyses were calculated as a supplement to the analyses described above and a summary is provided online. The original goal of this project was to calculate potential reactivity to item order through analyses that would be accessible to most researchers using questionnaires in their research. MGCFA requires not only specialized knowledge, but also specific software and the associated learning curve. We used *R* in our analyses, however, all analyses presented can be recreated with free software. The writers of *BayesFactor* have published online calculators for their work at http://pcl.missouri.edu/bayesfactor, and BF values are also available in *JASP* (JASP Team, 2018). The TOST analyses may be calculated using an Excel spreadsheet available from the author at https://osf.io/qzjaj/ or as an add-in module in the program *jamovi* (Jamovi project, 2018). Both *JASP* and *jamovi* are user friendly programs that researchers familiar with point and click software like Excel or SPSS will be able to use with ease.

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, as Tabachnick and Fidell (2012) have suggested that 5% or less of missing data may be safely filled in with minimal effects on hypothesis testing. Table 1 indicates the number of participants who were excluded for each set as a function of: 1) missing more than 5% of their data, 2) were missing data due to experimenter error (i.e., some versions of the PIL did not have one item, and these were excluded), or 3) missing values for the LPQ include participants who did not see this scale in some original rounds of the survey. Because we were examining context item-order effects, it did not seem prudent to include participants who were missing larger portions of their data, as it would be unclear if their context was the same as participants who did complete the entire survey. Our final sample sizes, as shown in Table 1 remained sufficiently large for analyses described below.

For participants with less than 5% missing data, we used the *mice* package in *R* to impute multiple datasets with those points filled in (Van Buuren & Groothuis-Oudshoorn, 2011). For the PIL randomized, $n = 43$ data points were imputed, $n = 60$ for the nonrandomized PIL, $n = 15$ for the randomized LPQ, and $n = 33$ for the nonrandomized LPQ. The advantage to using the *mice* package is the automatic estimation of missing data points based on the data type (i.e., 1-7 versus binary), rather than simple mean estimation. The default number of imputations is five, and one was selected to combine with the original dataset for analyses described below.

Next, each dataset was examined for multivariate outliers using Mahalanobis distance. As described in Tabachnick and Fidell (2012), Mahalanobis values were calculated for each participant based on their answer choice patterns for each of the twenty questions. These D values are compared to a $\chi^2(20)_{p<.001} = 45.31$, and observations with D values greater than this score were counted as outliers. This analysis is similar to using a z-score criterion of three standard deviations away from the mean. 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 and final sample size in each step is presented in Table 1.

PIL Analyses

Correlation Matrices. The correlation matrices for the randomized and nonrandomized versions of the PIL were found to be significantly different, $\chi^2(380) = 784.84$, $p < .001$. The Z score differences were examined, and 32 correlations were different across the possible 190 tests. A summary of differences can be found in Table 2. For each item, the total number of differences was calculated, as shown in column two, and those specific items are listed in column three. The last two columns summarize the directions of these effects. Positive Z-scores indicated stronger correlations between nonrandomized items, while negative Z-scores indicated stronger correlations for randomized items (summarized in the last column). Two items had strong context effects (i.e., impacted many items), item 2 *exciting life* and item 15 *prepared for death*. Interestingly, the impact is the reverse for these two items, as item 2 showed stronger relationships to items when randomized, while item 15 showed stronger relationships to items when nonrandomized.

Factor Loadings. Table 3 includes the factor loadings from the one-factor EFA. These loadings were compared using a dependent t-test matched on item, and they were not significantly different, $M_d = 0.00$, 95% CI $[-0.02, 0.03]$, $t(19) = 0.25$, $p = .802$. The effect size for this test was correspondingly negligible, d_{av} = -0.02 95% CI [-0.45, 0.42]. The TOST analysis 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, which indicated support for the null model.

Item Means. Table 3 includes the means and standard deviations 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 $[-0.13, -0.02]$, $t(19) = -2.91$, $p = .009$. 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 analysis indicated that the difference was within the range of a 5% percent change in item means (0.30). The TOST analysis for lower bound, $t(19)$ = -1.57, $p < .001$ and the upper bound, $t(19)$ = -4.26, $p < .001$, 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, 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 $M =$ 103.01 ($SD =$ 18.29) with excellent reliability ($\alpha =$ .93), while the mean for randomizing testing was $M =$ 104.48 ($SD =$ 17.81) with excellent reliability ($\alpha =$ .92). The total score difference was examined with an independent t-test and was not significant, $t(1,896) = -1.76$, $p =$ .079. The effect size for this difference was negligible, $d_s =$ -0.08 95% CI [-0.17, 0.29]. We tested if scores were changed by 5% (6.00 points), and the TOST analysis indicated that the lower, t(1897) = 5.43, $p <$ .001 and the upper bound, t(1897) = -8.95, $p <$ .001 were within this area of null change. The BF results also supported the null model, 0.25.

LPQ Analyses

Correlation Matrices. Mirroring the results for the PIL, the correlation matrices for the randomized and nonrandomized versions of the LPQ were significantly different, $\chi^2(380) = 681.72$, $p <$ .001. Less differences in correlation were found as compared to the PIL, only 19 out of the possible 190 combinations. The differences are summarized in Table 4. Most of the items affected one to four other items with item 13 *reliable person* showing the largest number of differences in correlation. All these changes were positive, meaning the correlations were larger for nonrandomized versions.

Factor Loadings. Table 5 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 $[-0.02$, 0.04], $t(19) = 0.97$, $p =$ .344. The difference found for this test was negligible, d_{av} = -0.07 95% CI [-0.50, 0.37]. The TOST analysis examined if any change was within .10 change, as described earlier. The lower, t(19) = -0.52, $p <$ .001 and the upper bound, t(19) = -1.42, $p <$ .001 were both significant, indicating that the found change was within the expected change. Further, in support of the null model, the BF was 0.34.

Item Means. Means and standard deviations of each item are presented in Table 5. 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 $[-0.02, 0.02]$, $t(19) = 0.26$, $p = .797$, 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 analysis indicated both lower, $t(19) = 0.48$, $p < .001$ and the upper bound, $t(19) = 0.04$, $p < .001$, were within the null range. The BF also supported the null model, 0.24.

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), with good reliability (α = .82), and the mean for nonrandomized testing was M = 14.19 (SD = 4.22) and good reliability (α = .84). An independent t-test indicated that testing did not change the total score, $t(1, 630) = 0.23$, $p = .819$. The effect size for this difference was negligible, d_s = 0.01 95% CI [-0.09, 0.45]. The TOST analysis indicated that the scores were within a 5% (1.00 points) change, lower: t(1627) = 5.13, p < .001 and upper: t(1627) = -4.67, p < .001. The BF results were in support of the null model, 0.06.

Discussion

As technology has advanced, initial research questioned the validity of online assessments versus paper assessments. With further investigation, several researchers discovered equivalence with regard to computer surveys compared with paper surveys (Deutskens et al., 2006; Lewis et al., 2009). However, with the addition of technology, Fang et al. (2012a) suggested that individuals respond with more extreme scores in online surveys than in-person surveys due to the social-desirability effect. Research on equivalence 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 are similar to those of Knowles et al. (1992), in that we found differences in correlation matrices when items were randomized versus nonrandomized. These differences may be attributed to the context of the items when randomized, as described by Salancik and Brand (1992). When viewed through a meaning-change (Knowles et al., 1992) or integration model (Rogers, 1974; Tourangeau & Rasinski, 1988), these differences may indicate that the context and background knowledge are shifting based on the order of the items presented.

As items showed these order context effects, randomization may present a way to combat those effects where the context of items is equalized across participants. However, it is important to show that randomization does not change the relationship of items with that underlying factor, rather just the context in which these items are presented. In both the PIL and LPQ scales, the factor loadings were found to be equivalent with results supporting the null hypothesis. For the PIL, we did find support for differences in item means using *p*-value criterion and Bayes Factor analyses. However, the effect size was small, meaning the differences were potentially 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. Potentially, the small difference in item means was due to fluctuating context and order effects, with more change possible using a 1 to 7 item answer format (i.e., more possible range of answer change). The LPQ item means were not found to differ, and the correlational analysis showed less items changed in contrast to the PIL analysis. 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. The match between results for two types of answer methodologies 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 and context effects described by the meaning-change hypotheses. Subject matter experts are usually involved in the scale development and this facet of reactivity should be considered in item development and deployment. 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. Randomization will create a wider range of possible interpretation-integration context scenarios as participants react and respond to items. This procedure would even out context effects at the sample or group level, but individual differences will be present for each participant.

Since we found equivalence in terms of overall scoring of the PIL and LPQ, we advise that randomization 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 (Henrich, Heine, & Norenzayan, 2010). As Fang et al. (2012b)'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 (Charters, 2004).

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 factor analyses can additionally be used to analyze a very similar set of questions (Brown, 2006); 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 (Valentine et al., 2017).

References

Aust, F., & Barth, M. (2017). papaja: Create APA manuscripts with R Markdown. Retrieved from https://github.com/crsh/papaja

Bargh, J. A., & Pratto, F. (1986). Individual construct accessibility and perceptual selection. *Journal of Experimental Social Psychology*, *22*(4), 293–311. doi:10.1016/0022-1031(86)90016-8

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. doi:10.1038/s41562-017-0189-z

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, *78*(2), 161–188. doi:10.1111/j.1751-5823.2010.00112.x

Brown, T. (2006). *Confirmatory factor analysis for applied research* (First Ed.). New York, NY: The Guilford Press.

Buchanan, E. M., Valentine, K. D., & Schulenberg, S. E. (2014). Exploratory and confirmatory factor analysis: Developing the Purpose in Life Test–Short Form. In P. Bindle (Ed.), *SAGE research methods cases.* London, UK: SAGE Publications, Ltd. doi:10.4135/978144627305013517794

Buchanan, E. M., Valentine, K. D., & Scofield, J. E. (2017). MOTE. Retrieved from https://github.com/doomlab/MOTE

Buchanan, T., Ali, T., Heffernan, T., Ling, J., Parrott, A., Rodgers, J., & Scholey, A. (2005). Nonequivalence of on-line and paper-and-pencil psychological tests: The case of the prospective memory questionnaire. *Behavior Research Methods*, *37*(1), 148–154. doi:10.3758/BF03206409

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Cantrell, M. A., & Lupinacci, P. (2007). Methodological issues in online data collection.

Journal of Advanced Nursing, *60*(5), 544–549. doi:10.1111/j.1365-2648.2007.04448.x

Charters, E. (2004). New perspectives on popular culture, science and technology: Web browsers and the new illiteracy. *College Quarterly*, *7*(1), 1–13.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second Ed.). Hillsdale, NJ: Earlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. doi:10.1037/0003-066X.49.12.997

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (Second Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in Web- or Internet-based surveys. *Educational and Psychological Measurement*, *60*(6), 821–836. doi:10.1177/00131640021970934

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, *60*(1), 1–10. doi:10.1002/jclp.10217

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of Web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 177–180. doi:10.3758/BF03195440

Crumbaugh, J. C., & Maholick, L. T. (1964). An experimental study in existentialism: The psychometric approach to Frankl's concept of noogenic neurosis. *Journal of Clinical Psychology*, *20*(2), 200–207. doi:10.1002/1097-4679(196404)20:2<200::AID-JCLP2270200203>3.0.CO;2-U

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals,* *and meta-analysis* (First Ed.). New York, NY: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi:10.1177/0956797613504966

De Leeuw, E. D., & Hox, J. J. (1988). The effects of response-stimulating factors on response rates and data quality in mail surveys: A test of Dillman's total design method. *Journal of Official Statistics*, *4*(3), 241–249.

Deutskens, E., de Ruyter, K., & Wetzels, M. (2006). An assessment of equivalence between online and mail surveys in service research. *Journal of Service Research*, *8*(4), 346–355. doi:10.1177/1094670506286323

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Fourth Ed.). Sage.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(July), 1–17. doi:10.3389/fpsyg.2014.00781

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2008). *Internet, mail, and mixed-mode surveys: The tailored design method* (Third Ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*(2), 170–177. doi:10.1037/1082-989X.1.2.170

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes Factor hypothesis test. *Statistical Science*, *32*(2), 313–329. doi:10.1214/16-STS599

Fang, J., Wen, C., & Pavur, R. (2012a). Participation willingness in web surveys: Exploring effect of sponsoring corporation's and survey provider's reputation. *Cyberpsychology, Behavior, and Social Networking*, *15*(4), 195–199. doi:10.1089/cyber.2011.0411

Fang, J., Wen, C., & Prybutok, V. R. (2012b). An assessment of equivalence between Internet and paper-based surveys: evidence from collectivistic cultures. *Quality & Quantity*, *48*(1), 493–506. doi:10.1007/s11135-012-9783-3

Feldman, J. M., & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, *73*(3), 421–435. doi:10.1037/0021-9010.73.3.421

Frick, A., Bächtiger, M. T., & Reips, U.-D. (2001). Financial incentives, personal information and dropout in online studies. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (First Ed., pp. 209–219). Lengerich, Germany: Pabst Science Publishers.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–53. doi:10.1037/a0015251

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi:10.1017/S0140525X0999152X

Higgins, E., & Lurie, L. (1983). Context, categorization, and recall: The "change-of-standard" effect. *Cognitive Psychology*, *15*(4), 525–547. doi:10.1016/0010-0285(83)90018-X

Hox, J. J., & De Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity*, *28*(4), 329–344. doi:10.1007/BF01097014

Hutzell, R. (1988). A review of the Purpose in Life Test. *International Forum for Logotherapy*, *11*(2), 89–101.

Ilieva, J., Baron, S., & Healy, N. M. (2002). On-line surveys in international marketing research: Pros and cons. *International Journal of Market Research*, *44*(3), 361–376.

Jamovi project. (2018). jamovi (Version 0.8)[Computer software]. Retrieved from https://www.jamovi.org

JASP Team. (2018). JASP (Version 0.8.6)[Computer software]. Retrieved from https://jasp-stats.org/

Joinson, A. (1999). Social desirability, anonymity, and Intemet-based questionnaires. *Behavior Research Methods, Instruments, & Computers*, *31*(3), 433–438. doi:10.3758/BF03200723

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical*

Association, *90*(430), 773–795. doi:10.2307/2291091

Keppel, G., & Wickens, T. (2004). *Design and analysis: A researcher's handbook* (Fourth Ed.). Upper Saddle River, NJ: Prentice Hall.

Knowles, E. S., Coker, M. C., Cook, D. A., Diercks, S. R., Irwin, M. E., Lundeen, E. J., . . . Sibicky, M. E. (1992). Order effects within personality measures. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (First Ed., pp. 221–236). New York, NY: Springer-Verlag.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00863

Lakens, D. (2017). Equivalence tests. *Social Psychological and Personality Science*, *8*(4), 355–362. doi:10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . . Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. doi:10.1038/s41562-018-0311-x

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course* (First Ed.). New York, NY: Cambridge University Press.

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology*, *61*(2), 107–116. doi:10.1080/00049530802105865

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243. doi:10.1037/0022-3514.47.6.1231

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of*

Mathematical Psychology, *72*, 19–32. doi:10.1016/J.JMP.2015.06.004

MacLeod, C., & Campbell, L. (1992). Memory accessibility and probability judgments: An experimental evaluation of the availability heuristic. *Journal of Personality and Social Psychology*, *63*(6), 890–902. doi:10.1037/0022-3514.63.6.890

Media. (2016). *The Total Audience Report: Q1 2016.*

Melton, A. M. A., & Schulenberg, S. E. (2008). On the measurement of meaning: Logotherapy's empirical contributions to humanistic psychology. *The Humanistic Psychologist*, *36*(1), 31–44. doi:10.1080/08873260701828870

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/BF02294825

Meyerson, P., & Tryon, W. W. (2003). Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 614–620. doi:10.3758/BF03195541

Morey, R. D. (2015). On verbal categories for the interpretation of Bayes factors. Retrieved from http://bayesfactor.blogspot.com/2015/01/on-verbal-categories-for-interpretation.html

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for common designs. Retrieved from https://cran.r-project.org/package=BayesFactor

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (First., pp. 61–87). Elsevier. doi:10.1016/B978-012099980-4/50004-6

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-Research: Ethics, security, design, and control in psychological research on the Internet. *Journal of Social Issues*, *58*(1), 161–176. doi:10.1111/1540-4560.00254

Olson, K. (2010). An examination of questionnaire evaluation by expert reviewers. *Field Methods*, *22*(4), 295–318. doi:10.1177/1525822X10379795

Panter, A. T., Tanaka, J. S., & Wellens, T. R. (1992). Psychometrics of order effects. In N.

Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (First Ed., pp. 249–264). New York, NY: Springer-Verlag.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change* (First Ed.). New York, NY: Springer-Verlag.

Posner, M. I. (1978). *Chronometric explorations of mind* (First Ed.). Hillsdale, NJ: Erlbaum.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2*(1), 13–43. doi:10.1207/S15328031US0201_02

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. doi:10.1026/1618-3169.49.4.243

Reips, U.-D. (2012). Using the Internet to collect data. In *APA handbook of research methods in psychology, vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* (Vol. 2, pp. 291–310). Washington, DC: American Psychological Association. doi:10.1037/13620-017

Revelle, W. (2017). psych: Procedures for psychological, psychometric, and personality research. Evanston, IL: Northwestern University. Retrieved from https://cran.r-project.org/package=psych

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. doi:10.1037/0033-2909.113.3.553

Rogers, T. (1974). An analysis of the stages underlying the process of responding to personality items. *Acta Psychologica*, *38*(3), 205–213. doi:10.1016/0001-6918(74)90034-1

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi:10.3758/PBR.16.2.225

Salancik, G. R., & Brand, J. F. (1992). Context influences on the meaning of work. In N. Schwarz & S. Sudman (Eds.), *Context efects in social and psychological research* (pp.

237–247). New York, NY: Springer-Verlag.

Sanou, B. (2017, July). ICT Facts and Figures 2017. Retrieved from http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf

Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. doi:10.1007/BF01068419

Schuldt, B. A., & Totten, J. W. (1994). Electronic mail vs. mail survey response rates. *Marketing Research*, *6*, 36–39.

Schulenberg, S. E. (2004). A psychometric investigation of logotherapy measures and the Outcome Questionnaire (OQ-45.2). *North American Journal of Psychology*, *6*(3), 477–492.

Schulenberg, S. E., & Melton, A. M. A. (2010). A confirmatory factor-analytic evaluation of the purpose in life test: Preliminary psychometric support for a replicable two-factor model. *Journal of Happiness Studies*, *11*(1), 95–111. doi:10.1007/s10902-008-9124-3

Schulenberg, S. E., & Yutrzenka, B. A. (1999). The equivalence of computerized and paper-and-pencil psychological instruments: Implications for measures of negative affect. *Behavior Research Methods, Instruments, & Computers*, *31*(2), 315–321. doi:10.3758/BF03207726

Schulenberg, S. E., & Yutrzenka, B. A. (2001). Equivalence of computerized and conventional versions of the Beck Depression Inventory-II (BDI-II). *Current Psychology*, *20*(3), 216–230. doi:10.1007/s12144-001-1008-1

Schulenberg, S. E., Schnetzer, L. W., & Buchanan, E. M. (2011). The Purpose in Life Test-Short Form: Development and psychometric support. *Journal of Happiness Studies*, *12*(5), 861–876. doi:10.1007/s10902-010-9231-9

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*(4), 605–632.

doi:10.1177/001316440121971392

Smyth, J. D. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly*, *70*(1), 66–77. doi:10.1093/poq/nfj007

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107. doi:10.1086/209528

Strack, F., & Martin, L. L. (1987). Thinking, judging, and communicating: A process account of context effects in attitude surveys. In *Recent research in psychology* (pp. 123–148). New York, NY: Springer. doi:10.1007/978-1-4612-4798-2_7

Strack, F., Schwarz, N., & Gschneidinger, E. (1985). Happiness and reminiscing: The role of time perspective, affect, and mode of thinking. *Journal of Personality and Social Psychology*, *49*(6), 1460–1469. doi:10.1037/0022-3514.49.6.1460

Strack, K. M., & Schulenberg, S. E. (2009). Understanding empowerment, meaning, and perceived coercion in individuals with serious mental illness. *Journal of Clinical Psychology*, *65*(10), 1137–1148. doi:10.1002/jclp.20607

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (Sixth Ed.). Boston, MA: Pearson.

Tesser, A. (1978). Self-generated attitude change. In *Advances in experimental social psychology* (Vol. 11, pp. 289–338). Elsevier. doi:10.1016/S0065-2601(08)60010-6

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299–314. doi:10.1037/0033-2909.103.3.299

Tourangeau, R., Rips, L. J., & Rasinski, K. (1999). *The psychology of survey response* (First Ed.). Cambridge, UK: Cambridge University Press.

Trent, L. R., Buchanan, E., Ebesutani, C., Ale, C. M., Heiden, L., Hight, T. L., … Young, J. (2013). A measurement invariance examination of the Revised Child Anxiety and Depression Scale in a southern sample: Differential item functioning between African

American and Caucasian youth. *Assessment*, *20*(2), 175–187. doi:10.1177/1073191112450907

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. doi:10.1016/0010-0285(73)90033-9

Valentine, K. D., Buchanan, E. M., Scofield, J. E., & Beauchamp, M. (2017). Beyond p-values: Utilizing multiple estimates to evaluate evidence, 1–29. doi:10.17605/osf.io/9hp7y

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. doi:10.18637/jss.v045.i03

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi:10.3758/BF03194105

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. doi:10.1177/0963721416643289

Webb, E. S., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). *Unobtrusive measures: Nonreactive research in the social sciences* (First Ed.). Chicago, IL: Rand McNally.

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods*, *18*(1), 53–70. doi:10.1037/a0031607

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. doi:10.1177/0011000006288127

Table 1

Demographic and Data Screening Information

Group	Female	White	Age (SD)	Original N	Missing N	Outlier N	Final N
PIL Random	61.6	81.1	19.50 (2.93)	1462	333	59	1070
PIL Not Random	54.1	78.6	19.68 (3.58)	915	51	36	828
LPQ Random	-	-	-	1462	555	24	883
LPQ Not Random	-	-	-	915	150	16	749

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. Columns titled 'Female' and 'White' are represented by percentages.

Table 2

Correlation Matrices Results by Item for the PIL

Item	Differences	Items Changed	Direction of Change	Stronger Randomized
1	3	2, 12, 15	2 Negative; 1 Positive	2 & 12
2	9	1, 3, 4, 8, 9, 15, 18, 19, 20	8 Negative; 1 Positive	1, 3, 4, 8, 9, 18, 19, 20
3	1	2	1 Negative	2
4	2	2, 15	1 Negative; 1 Positive	2
5	2	9, 15	1 Negative; 1 Positive	9
6	2	12, 15	2 Positive	N/A
7	2	17, 19	2 Positive	N/A
8	1	2	1 Negative	2
9	3	2, 5, 15	2 Negative; 1 Positive	2 & 5
10	2	12, 15	2 Positive	N/A
11	3	12, 15, 20	3 Positive	N/A
12	6	1, 6, 10, 11, 14, 20	2 Negative; 4 Positive	1 & 14
13	0	N/A	N/A	N/A
14	2	12, 18	2 Negative	12 & 18
15	10	1, 2, 4, 5, 6, 9, 10, 11, 17, 19	10 Positive	N/A
16	0	N/A	N/A	N/A
17	4	7, 15, 18, 19	4 Positive	N/A
18	3	2, 14, 17	2 Negative; 1 Positive	2 & 14
19	5	2, 7, 15, 17, 20	1 Negative; 4 Positive	2
20	4	2, 11, 12, 19	1 Negative; 3 Positive	2

Table 3

Item Statistics for the PIL Scale

Item	FL-R	FL-NR	M-R	M-NR	SD-R	SD-NR
1	.667	.638	4.829	4.806	1.279	1.278
2	.679	.572	4.929	4.600	1.437	1.452
3	.685	.671	5.815	5.732	1.124	1.101
4	.839	.847	5.673	5.655	1.300	1.285
5	.639	.574	4.666	4.407	1.496	1.497
6	.674	.685	5.425	5.338	1.308	1.400
7	.424	.439	6.172	6.081	1.207	1.373
8	.626	.596	5.014	5.011	1.092	1.139
9	.823	.796	5.355	5.327	1.176	1.198
10	.723	.764	5.202	5.156	1.502	1.543
11	.775	.796	5.222	5.165	1.629	1.621
12	.604	.649	4.496	4.527	1.570	1.600
13	.429	.403	5.745	5.738	1.244	1.216
14	.449	.421	5.431	5.239	1.377	1.547
15	.081	.211	4.376	4.149	1.941	1.884
16	.547	.554	5.099	5.266	1.983	1.861
17	.720	.735	5.422	5.399	1.393	1.404
18	.483	.501	5.387	5.302	1.474	1.593
19	.678	.721	4.879	4.907	1.412	1.455
20	.782	.810	5.343	5.210	1.314	1.289

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

Table 4

Correlation Matrices Results by Item for the LPQ

Item	Differences	Items Changed	Direction of Change	Stronger Randomized
1	3	11, 13, 18	1 Negative; 2 Positive	18
2	1	6	1 Positive	N/A
3	1	8	1 Negative	8
4	0	N/A	N/A	N/A
5	2	6, 11	2 Positive	N/A
6	2	2, 5	2 Positive	N/A
7	3	13, 18, 20	3 Positive	N/A
8	2	3, 20	2 Negative	3 & 20
9	2	11, 13	2 Positive	N/A
10	1	20	1 Positive	N/A
11	4	1, 5, 9, 12	4 Positive	N/A
12	2	11, 13	2 Positive	N/A
13	6	1, 7, 12, 15, 16	6 Positive	N/A
14	0	N/A	N/A	N/A
15	0	N/A	N/A	N/A
16	1	13	1 Positive	N/A
17	1	13	1 Positive	N/A
18	3	1, 7, 20	1 Negative; 2 Positive	1
19	0	N/A	N/A	N/A
20	4	7, 8, 10, 18	1 Negative; 3 Positive	8

Table 5

Item Statistics for the LPQ Scale

Item	FL-R	FL-NR	M-R	M-NR	SD-R	SD-NR
1	.675	.682	.567	.613	.496	.487
2	.900	.870	.754	.760	.431	.428
3	.503	.394	.864	.844	.343	.363
4	.730	.685	.908	.868	.289	.339
5	.687	.682	.419	.507	.494	.500
6	.502	.555	.638	.582	.481	.494
7	.193	.286	.775	.810	.418	.392
8	.555	.471	.482	.467	.500	.499
9	.856	.911	.810	.781	.393	.414
10	.592	.620	.635	.646	.482	.478
11	.636	.760	.727	.761	.446	.427
12	.687	.758	.787	.752	.410	.432
13	.314	.399	.965	.911	.184	.286
14	.486	.486	.762	.769	.426	.422
15	.046	.102	.323	.395	.468	.489
16	.700	.707	.863	.872	.344	.335
17	.514	.502	.847	.814	.360	.389
18	.558	.511	.830	.828	.376	.378
19	.675	.717	.463	.497	.499	.500
20	.644	.618	.721	.712	.449	.453

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