AIS Transactions on Human-Computer Interaction

Volume 12 | Issue 2 Article 1

6-30-2020

Do We Truly Sacrifice Truth for Simplicity: Comparing Complete Individual Randomization and Semi-Randomized Approaches to Survey Administration

Eleanor Loiacono
WPI, eloiacon@wpi.edu

Vance Wilson
Worcester Polytechnic Institute, vancewilson@gmail.com

Recommended Citation

Loiacono, E., & Wilson, V. (2020). Do We Truly Sacrifice Truth for Simplicity: Comparing Complete Individual Randomization and Semi-Randomized Approaches to Survey Administration. *AIS Transactions on Human-Computer Interaction, 12*(2), 45-69. https://doi.org/10.17705/1thci.00128
DOI: 10.17705/1thci.00128

Transactions on Human-Computer Interaction

Volume 12 Issue 2

6-2020

Do We Truly Sacrifice Truth for Simplicity: Comparing Complete Individual Randomization and Semi-Randomized Approaches to Survey Administration

Eleanor Loiacono

Worcester Polytechnic Institute, eloiacon@wpi.edu

E. Vance Wilson

Worcester Polytechnic Institute, vancewilson@gmail.com

Recommended Citation

Loiacono, E., & Wilson, E. V. (2020). Do we truly sacrifice truth for simplicity: Comparing complete individual randomization and semi-randomized approaches to survey administration. *AIS Transactions on Human-Computer Interaction*, *12*(2), 45-69.

DOI: 10.17705/1thci.00128

Available at http://aisel.aisnet.org/thci/vol12/iss2/1

Transactions on Human-Computer Interaction

Research Paper DOI: 10.17705/1thci.00128 ISSN: 1944-3900

Do We Truly Sacrifice Truth for Simplicity: Comparing Complete Individual Randomization and Semi-randomized Approaches to Survey Administration

Eleanor Loiacono
Worcester Polytechnic Institute

E. Vance Wilson
Worcester Polytechnic Institute

Abstract:

Human-computer interaction researchers have long used survey methodologies. However, debate remains about the potential for participants to provide biased responses to subsequent items based on previously viewed items. In this research, we investigate the effects of survey item ordering that researchers have not studied previously. Grounded in previous exploratory item-ordering studies using an HCI online survey, we investigate bias in more detail. In addition, we use an adult sample population so that we can extend our results more broadly as compared to previous research. We employed two distinct randomizing survey approaches: 1) complete item randomization for each respondent (random), which presents items to each respondent in a completely randomized order; and 2) partially individualized item randomization (grouped), which presents constructs in the same order in a survey but randomizes items in each construct for each respondent. Our results suggest researchers should use fully randomized survey instruments in HCI research whenever possible since grouped ordering of any kind increases bias and statistical inflation, which can influence results' veracity. Additionally, we did not appear to find any significant increase in the participants' frustration or fatigue to be associated with the random treatment.

Keywords: Survey Administration, Item Ordering, Reliability, Item Randomization.

Gregory D. Moody was the accepting senior editor for this paper.

1 Introduction

Human-computer interaction (HCI) research relies heavily on survey design to define and measure relevant constructs. HCI research has used contemporary methods in instrument development. However, researchers continue to change and improve techniques to help create better and more accurate measures. In reviewing the *Association for Information Systems Transactions on Human Computer Interactions* (*AIS THCI*), which senior scholars recognize as an especially relevant journal, we found nearly 35 percent of papers in the journal used the survey methodology as their basis. Given this finding, it seems imperative that HCI researchers continually look to improve on existing approaches, based on new and developing techniques, to foster and maintain best practices (MacKenzie, Podsakoff, & Podsakoff, 2011).

Researchers who develop and employ surveys focus on improving measurement and reducing the potential for method bias. As recent literature highlights, method bias can cause significant negative impacts on survey results (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Podsakoff, MacKenzie, & Podsakoff, 2012) via introducing systematic error. Such errors can result in flawed instruments, inaccurate conclusions, and poor theory development. Thus, researchers must continually strive to decrease the influence of method bias through practices such as greater item randomization. Though some IS studies have looked into item ordering and its influence on survey results, they have 1) included only college student respondents and 2) focused on broad comparisons between individually randomized item-ordering surveys and surveys that administered items in static construct groupings. We go much farther in this study.

In the past, perfect individual randomization was impractical. Paper and pencil-based surveys made it difficult to create a fully randomized survey for each participant. Such randomization would make administrating and coding data difficult. Today, however, technological advances in online survey administration make such randomization simple and easy. For example, Qualtrics, SurveyMonkey, and SurveyGizmo all provide randomized question capabilities.

Though method bias poses a serious concern to HCI researchers and one can easily operationalize randomized items for individual participants today, only a few studies in quantitative psychology (Buchanon et al., 2018) and information systems (Wilson & Lankton, 2012; Wilson, Srite, & Loiacono, 2017) have investigated the influence that item ordering (individually randomized versus grouped) has on method bias. In this paper, we look at two distinct item-randomizing approaches to determine the impact item grouping can have on survey results: 1) complete item randomization, which presents items to each respondent (random) in a completely randomized order; and 2) partially individualized item randomization (grouped), which presents constructs in the same order in a survey but randomizes items in each construct for each respondent.

2 Background

Virtually the entire early literature on item ordering has used static surveys that researchers have administered using pencil-and-paper methods (Tourangeau, Rips, & Rasinski, 2000). Researchers have long investigated the potential for participants to provide biased responses to subsequent items based on items they viewed earlier in the same survey, a phenomenon called a context effect (Schwarz & Sudman, 2012) that arises primarily through semantic and sequential mechanisms (Krosnick & Presser, 2010). Tourangeau et al. (2000, p. 197) explain that:

> *The judgments called for by attitude questions are rarely absolute but are typically made in relation to some standard, generally an implicit one. It is hardly surprising, then, that attitude judgments turn out to be quite context-dependent. As survey researchers have demonstrated repeatedly, the same question often produces quite different answers, depending on the context.*

Decades ago, researchers reported that significant tradeoffs exist in choosing among static ordering schemes in attempting to avoid systematic bias that arises from context effects (Harrison & McLaughlin, 1993; Kraut, Wolfson, & Rothenberg, 1975; Schuman & Presser, 1981). Yet, identifying the most appropriate tradeoff proved contentious. Historically, two camps arose among researchers regarding the preferred form for ordering items in the area's static survey designs: one that argued for grouped static ordering and another that argued for intermixed static ordering (Budd, 1987; Schriesheim, Solomon, Kopelman, 1989).

Two IS research papers illustrate these arguments' central tenets. Davis and Venkatesh (1996) conducted three experiments among undergraduate students to study the perceived usefulness and ease of use constructs from the technology acceptance model (Davis, 1989). They compared grouped static ordering

with two versions of intermixed static ordering: 1) a version that alternated items' order between the two constructs and 2) a version that used randomized ordering. They reported that neither intermixing form significantly changed the two constructs' reliability or validity and concluded that "by ruling out carryover biases in these scales, this study is encouraging to practitioners who use these scales to evaluate risky investments in information technologies" (Davis & Venkatesh, 1996, p. 42).

Subsequently, Goodhue and Loiacono (2002) administered the WebQual survey instrument (Loiacono, Watson, & Goodhue, 2002, 2007) to undergraduate students in two versions: a grouped static version in which they labeled sections with the construct name and an intermixed static version in which they randomized items' order. They found Cronbach's alpha reliability for constructs in the grouped static version to be significantly higher than in the intermixed static version. As such, they concluded that "Cronbach's alpha is artificially inflated when questions are placed adjacent to each other and labeled, and that intermixing questions results in a small but systematic improvement in actual reliability" (Goodhue & Loiacono, 2002, p. 9).

Because both camps in the grouped versus intermixed survey argument could point to the benefits from their preferred approach, their argument had no clear winner in the context of IS survey research or, for that matter, in research conducted in other fields (Tourangeau et al., 2000). As a result, IS researchers have continued to approach decisions about ordering items in surveys based largely on personal preferences and prejudices rather than clear guidelines.

2.1 Advent of Individual Randomization via Online Surveys

In the intervening period, important structural changes have occurred in the survey research landscape. Whereas Davis and Venkatesh (1996) and Goodhue and Loiacono (2002) studied pencil-and-paper designs, IS researchers now increasingly administer surveys in IS research online where, rather than intermixing items in one or a few versions, they can programmatically randomize items' order for each participant without adding to survey administrators' or data analysts' workload.

Many major commercial online survey applications such as Qualtrics, SurveyMonkey, and SurveyGizmo allow one to individually randomize the order in which items appear. SurveyGizmo (2015) describes its "RandomizeQuestions" capability in this way: "Survey researchers frequently use randomization as a tool to combat survey bias. Randomizing the order of questions, pages, and/or answer options in your survey prevents bias introduced by order and/or survey fatigue.". Despite this explanation's logic, it principally relies on anecdotal evidence. Although literature has examined the effects that item randomization has on intermixed static survey designs, we found little empirical research that compares individually randomized designs to static designs despite several papers that promote benefits of using individual randomization (Bradlow & Fitzsimons, 2001; Raghunathan & Grizzle, 1995; Weinberger, Darkes, Del Boca, Greenbaum, & Goldman 2006).

In this paper, we investigate effects of survey item ordering that researchers have not studied previously but that can have a significant impact on HCI research design. We ground our approach in research by Wilson and Lankton (2012) and Wilson et al. (2017), who conducted exploratory item-ordering studies using an IS online survey. Along with providing insights regarding empirical effects of individually randomized item ordering, they identified several important issues that researchers have scarcely addressed, such as unexpectedly high levels of item response anomalies they encountered in their study's grouped-static condition. However, these studies had two important limitations. First, they used student populations, which limited their generalizability to the broader adult population. Second, they focused on broad comparisons between individually randomized item-ordering surveys and surveys that administered items in static construct groupings.

Our research design contributes to the literature in several other meaningful ways. First, we study a different survey instrument than the ones that Wilson and Lankton (2012) and Wilson et al. (2017) used. Thus, we extend the validity of these previous findings. Second, we include a broad spectrum of adult respondents so we can better extend the results beyond previous research. Further, we examine the effects of two distinct randomizing survey approaches that online survey software makes possible.

The random condition provides an experimental control by ensuring that we eliminate systematic semantic and sequential context effects from the survey results. However, researchers have argued that survey items need to maintain a coherent flow in order for participants to understand them and to decrease how much frustration and fatigue they experience (Davis & Venkatesh, 1996; Krosnick & Presser, 2010). The grouped

condition specifically orders constructs and individually randomizes items' order in each construct and, thereby, maintains coherent ordering.

Understanding tradeoffs between item-ordering strategies can be important to HCI survey researchers. Previous research has suggested that statically grouping items in constructs falsely artificially inflates construct reliability and validity statistics (Goodhue & Loiacono, 2002; Wilson & Lankton, 2012; Wilson et al., 2017) and increases common method bias (Straub, Boudreau, & Gefen, 2004). Further, the resulting structural models developed from statically grouped data may suffer from lower model fit and less parsimony than models developed using individually randomized data (Wilson et al., 2017).

We contribute broadly to the HCI field. First, we hypothesize and test the impact of complete item randomization versus partial item grouping, which will help researchers better understand the tradeoffs between item-ordering strategies. In terms of generalizability, by using a broad spectrum of adult respondents, we can extend our results beyond previous research, which has used data solely from university students. Further, by assessing semantic and sequential effects in these survey administrations, we help researchers understand the consequences that arise from using random versus partially grouped item surveys. Finally, we provide the hypothetical effects that randomization has on instrument development, which researchers should consider when developing multi-item measures. We highlight past studies on item ordering and their methods in Table 1.

3 Literature Review and Hypotheses Development

According to the classical test theory (CTT), psychometric measures do not produce true scores (Lord & Novick, 1968); rather, they produce approximations. These observed scores (X) are the sum of the true (error free) measure (T) plus random error (e) (i.e., X = T + e).

As Kane (2008) notes, a "true score is a construct that we can imagine, talk about and estimate, but we cannot observe it directly. We define it, or construct it, as the expected value over all possible observed scores, because this choice is conceptually useful" (p. 9). Random errors are just that—random. They are unknown, unpredictable influences that one cannot easily replicate by repeating an experiment (Taylor, 1999). Though researchers cannot avoid random error, they can account for it through statistical analysis and increased sample size. Systematic errors, on the other hand, are repeatable errors that emerge due to a flawed experimental design and usually directional (Taylor, 1999). They can lead to serious problems in interpreting the results:

> *Because of resulting biased measurement, systematic error can lead to seriously misleading conclusions and in particular such decisions based on obtained test scores. Hence, systematic measurement error must be eliminated, or informed arguments in favor of it not being present in tests under consideration must be justified, before one can place trust in given instruments and recommend them for use. In fact, lack of systematic error of measurement is an essential requirement for any measurement instrument. Without this requirement being fulfilled, it cannot be claimed that a particular instrument is a really trustworthy device of behavioral measurement.* (Raykov & Marcoulides, 2010, p.116)

Though no study can be completely free from error, researchers have a duty to attempt to reduce how much systematic error their studies contain as much as possible. In order to reduce systematic error when administering surveys, researchers must carefully consider the way in which they communicate questions to respondents. For example, participants responding to later questions in a survey may consider responses they made to earlier questions. Additionally, later questions in a survey may suffer from respondent fatigue. These examples show just some ways that "question wording, form, and context can lead to measurement error" (Lavrakas, 2008).

Measurement errors influence a measure's reliability and validity. For example, a biased survey instrument will show inconsistent scores—sometimes high and sometimes low—depending on its administration. Further, hypothesis guessing from similarly placed items in a survey may lead to the appearance of increased construct validity where it does not exist (McCroskey, Richmond, & McCroskey, 2006).

Table 1. Past Literature Studies and Methods

Study reference	Treatment*				Sample**	Method bias	Validity	Reliability	Fatigue	Frustration
	Indiv. random	Static random	Static group	Static mix						
Current paper	X		X		X	X	X	X	X	X
Buchanon et al. (2018)	X		X		S	X		X		
Wilson et al. (2017)	X		X		S	X	X	X		
Podsakoff et al. (2012)						X	X	X		
Wilson & Lankton (2012)	X		X		S	X	X	X		
Schwarz & Sudman (2012)						X				
Krosnick & Presser (2010)									X	X
Weinberger et al. (2006)		X			S		X			
Straub et al. (2004)						X	X	X		
Tourangeau, Couper, & Conrad (2004)					X	X				
Podsakoff et al. (2003)						X				
Goodhue & Loiacono (2002)		X	X		S	X	X	X	Confusion and annoyance	
Bradlow & Fitzsimons (2001)					S					
Davis & Venkatesh (1996)			X	X	S	X	X	X		
Raghunathan & Grizzle (1995)		X	X			X		X		
Harrison & McLaughlin (1993)		X	X		S & E			X		
Schriesheim et al. (1989)		X	X		S	X	X	X		
Tourangeau, & Rasinski (1988)						X				
Budd (1987)			X	X		X	X	X		
Schuman & Presser (1981)						X				
Kraut et al. (1975)				X					X	

* **Individual random** = items completely randomized for each respondent; **static random** = several versions of survey with items randomized differently in each; **static group** = items grouped on survey by construct but items are randomized in the group for each respondent; and **static mix** = items are grouped on survey by construct and remain in the same order for all respondents.
** S = student sample, E = employees, X = diverse adult sample.

More specifically, systematic error that can arise from using grouped versus random item ordering concerns the effect that such ordering can have on not only reliability and construct validity but also context validity, common method bias, and participant frustration. In this study, we focus on a particular type of reliability: a measure's internal consistency, which assesses results' stability across items in a survey (Hair, Black, Babin, & Anderson, 2010). If a survey is truly reliable, it will have highly consistent items such that items measuring the same underlying construct will receive similar scores. Though researchers cannot exactly calculate reliability, they often estimate it using Cronbach's alpha (Arnetz & Berg, 1993) with a recommended cutoff of 0.70 (Hair et al., 2010). Recently, however, researchers have cautioned against reporting on alpha reliabilities alone since SEM more accurately calculates reliability statistics using composite reliabilities (Green & Yang, 2009; Sijtsma, 2009). Thus, IS researchers recommend augmenting Cronbach's alpha with composite reliability statistics (e.g., Hess, McNab, & Basoglu, 2014; Lowry, Gaskin, & Moody, 2015).

Previous research has shown that the way one places items in a survey can affect reliability (Brod, 1984). If one places all items that relate to an underlying construct near each other, then participants will likely respond similarly to all items simply to avoid cognitive dissonance (Lazarus, 1966). Additionally, participants may, by seeing items in proximity to each other, guess or extrapolate the expected relationship and conform their answers to it (Lazarus, 1993). In such cases, the item placement artificially inflates the resulting reliabilities.

Research that has examined item ordering has had mixed results. Budd (1987) found that grouped items increased reliability and path coefficients. On the other hand, Davis and Venkatesh (1996) did not find an increase in reliabilities, validities, or path coefficients. In fact, they found that their randomized item ordering actually caused more confusion and frustration in their participants. Such confusion, they believed, could introduce error into the process and, thus, decrease reliability measures.

Goodhue and Loiacono (2002) employed a newly developed measure of website quality, WebQual, and found results that support Budd's (1987) results: grouping items artificially inflated Cronbach's alpha scores. More recent studies have found the same effect when comparing grouped versus individually randomized surveys as well (Wilson & Lankton, 2012; Wilson et al., 2017). Since grouped items tend to have correlated errors, which will not likely cancel each other out, they will tend to artificially increase reliability. Randomization, however, actually increases the "true" (more accurate) reliability of the measures (i.e., items with less correlated error terms) by showing an improvement in the correlations with a related construct (intent to reuse) when one uses a composite measure compared to when one employs an average of individual question correlations. Goodhue and Loiacono (2002) also found a small increase in the path coefficients but believed it to arise from confusion on the part of randomized treatment participants, who responded with less consistency than participants in the grouped treatment. Thus, we hypothesize that:

H1: Grouped surveys have inflated reliability compared to randomized surveys.

Construct validity ensures that a survey accurately measures what it should measure (Cronbach & Meehl, 1995). Construct validity has two major subcategories: convergent and discriminant validity. Convergent validity refers to the degree to which items that measure the same construct relate to one another. Conversely, discriminant validity examines whether items that measure different constructs do not relate to one another (Campbell, 1959). We can demonstrate an example using measures from the technology acceptance model (TAM). The fact that the items that measure ease of use correlate more with other items intended to measure ease of use indicates convergent validity, while the fact that these items correlate less with the items that measure usefulness demonstrates discriminant validity. Thus, we hypothesize that:

H2: Grouped surveys have inflated construct validity statistics compared to randomized surveys.

Common method bias refers to incorrect "variance that is attributable to the measurement method rather than to the constructs the measures are assumed to represent" (Podsakoff et al., 2003, p. 879) and creates a false internal consistency. Where common method bias exists, measures may have inflated or deflated correlations based on their common source (Williams & Brown, 1994). As Chang, van Witteloostuijn, and Eden (2010, p. 178) note:

> *[S]elf-report data can create false correlations if the respondents have a propensity to provide consistent answers to survey questions that are otherwise not related. Thus, common methods can cause systematic measurement errors that either inflate or deflate the observed relationships between constructs, generating both Type I and Type II errors.*

Several types of common method bias pertain to this research (for complete list, see Podsakoff et al., 2003). First, item context effects refer to "any influence or interpretation that a respondent might ascribe to an item solely because of its relation to the other items making up an instrument" (Podsakoff et al., 2003, p. 882; Wainer & Kiely, 1987). Second, the way one groups items or constructs in a questionnaire may decrease intraconstruct correlations and increase interconstruct correlations (Podsakoff et al., 2003). Thus, we hypothesize that:

H3: Grouped surveys have increased common method bias effects compared to randomized surveys.

Context effects occur when a prior item in a survey affects a respondent's answer to a later item in a survey. In a sense, it primes respondents to answer subsequent items in a similar way. For example, if respondents see a question that asks "was the website difficult to use" and they respond "difficult", then they will be more likely to answer the next related question they see in a similar manner, such as "did you find the website hard to use", to avoid cognitive dissonance or effort. This "biasing" in thinking and answering reduces a measure's accuracy. Context effects occur frequently and can have significant effects between proximate items (Weijters, Geuens, & Schillewaert, 2009). Semantic and sequential effects constitute two major context effects.

A semantic effect occurs as participants interpret what later items mean based on items to which they have previously responded, such as reporting heightened levels of criminal victimization following general questions about crime (Krosnick & Presser, 2010). Though logically grouping items can facilitate respondents' cognitive processing by making meanings clearer or memory retrieval easier, it can also introduce systematic error (Knowles, 1988; Knowles & Byers, 1996). However, others have found no to little error (Smith, 1983; Martin, 1980). Specifically, Couper, Traugott, and Lamias (2001) and Tourangeau et al. (2004) found in their Web survey studies that items correlated more when the authors presented the items together on a screen than when they presented the items individually on a screen. By randomizing ordering in constructs in the grouped conditions, we anticipate that we will find lower semantic context effects compared to the ones we find using static designs. However, researchers have established that context effects vary in size statistically as a function of proximity between items with the strongest effects occurring between adjacent items (Feldman & Lynch, 1988; Smith, 1988; Weijters et al., 2009); thus, we anticipate that we may find some level of semantic context effects. Thus, we hypothesize that:

H4: Semantic context effects occur in in the grouped survey format.

Sequential context effects occur due to one's placing items at a set position in a survey. For example, Kraut et al. (1975) found smaller standard deviations, less discriminating responses, and less frequent extreme responses when they moved items to a later survey position, and Knowles (1988) found that items positioned later in a survey correlated more strongly with the total score, which suggests increased consistency. In reviewing the literature, Krosnick and Presser (2010) observed that sequential order effects normatively accompany changes in participants' motivation, learning, and fatigue that occur as they progress through a survey. Items placed earlier in a survey may influence responses to later questions as respondents "learn" about the survey and their role (Krosnick & Presser, 2010). They may begin answering questions more accurately as they become more familiar with the survey. Several studies have also found that participant fatigue can result in more missing data, less detailed answers, and less differentiation among questions when placed later rather than earlier in a survey (Johnson, Sievesking, & Clanton, 1974; Kraut et al., 1975, Herzog & Bechman, 1991; Backor, Golde, & Nie, 2007). Such effects can result in the error variance increasing as time passes (Hess et al., 2012).

Previous research has also noted time-based effects. For example, Tourangeau and Rasinksi (1988) found that participants underreported the frequency with which they visited the dentist as the questions in the survey they took increased, and Bradley and Daly (1994) found that participants in a study on train service options provided erratic responses as the survey they took progressed. Additionally, Caussade, Ortúzar, Rizzi, and Hensher (2005) reported participant learning and fatigue when they asked participants to estimate the value of different travel options. Specific to IS-related research that examines the effects of individual randomization, Wilson and Lankton (2012) found systematic differences in subjects' responses between a survey's early and late stages. Thus, we hypothesize that:

H5: Sequential context effects occur in the grouped survey format.

The issues of respondent fatigue and frustration that Davis and Venkatesh (1996) raised have received little subsequent attention. Researchers need to consider these issues to avoid confusing and annoying

respondents and, thus, obtaining inaccurate data. Through open-ended verbal questioning, Davis and Venkatesh (1996) found that "subjects were more confused and annoyed when items were intermixed, suggesting a tendency toward ''output interference'' effects, which themselves could have a biasing effect" (p. 19). Thus, we hypothesize that:

H6: Grouped survey participants have lower fatigue and frustration compared to randomized survey participants.

4 Methodology

4.1 Participants

We collected data for this research from adult Internet users ranging from 18 to 81 years old. We used a marketing research firm to gather respondents in the targeted group (adult Internet users). We also collected additional demographic information, such as gender, education level, and age (see Table 2).

Table 2. Demographic Information

Gender	Male: 50% Female: 50%
Education level	61.5 had at least a college (bachelors) degree
Age	Mean = 36.9 years S.D. = 11.6 years Skewness = 0.92 Age distribution: 20s = 30.9% 30s = 38% 40 = 13.6% 50s = 12.3% 60s = 4.3% 70s = 0.4%

4.2 Measures

Similar to previous item randomization studies, we used an HCI-established instrument. We chose the 36-item WebQual survey (Loiacono, Watson, & Goodhue, 2007; Appendix A), which focuses on how users perceive website quality, for this study. It contains 12 first-order constructs: information fit-to-task, tailored communication, trust, response time, ease of understanding, intuitive operations, visual appeal, innovativeness, emotional appeal, consistent image, online completeness, and relative advantage. Additionally, it includes measures for website reuse and purchase intention.

To capture each participant's fatigue and frustration levels, we included measures consistent with Wilson and Lankton (2012) at the end of the survey. We also included an open-ended question that allowed participants to share their opinions about the survey at the end of the survey. We included these measures and question to determine fatigue and frustration levels between groups. We measured all items on a seven-point Likert scale (1 = "strongly disagree" and 7 = "strongly agree") (see Appendix A).

4.3 Survey Formats

We developed two survey formats to reflect the different test conditions: randomized and group randomized (which we refer to as "grouped" henceforth). Participants completed 506 surveys; however, we deleted 105 responses (shared almost equally between the two conditions) after we cleaned the data to ensure its quality. First, since the survey specifically asked participants to review a website for at least five minutes between answering the demographic questions and the rest of the survey, we removed 33 individuals who took less than five minutes to respond to the survey from the dataset. Next, we removed six participants who responded incorrectly to an attention-check question. Finally, we removed 65 participants who responded incorrectly to one of the three task questions. As a result, we had 402 useable responses.

Of the total sample, 200 completed the randomized survey and 202 completed the grouped survey. The randomized survey comprised one page of WebQual items that we fully randomized. The grouped treatment also comprised one page of WebQual items with constructs in a consistent order but randomly ordered items in each construct. Figure 1 depicts both survey treatments. We administered the survey using an online survey service, Qualtrics, with which we could create the two survey versions.

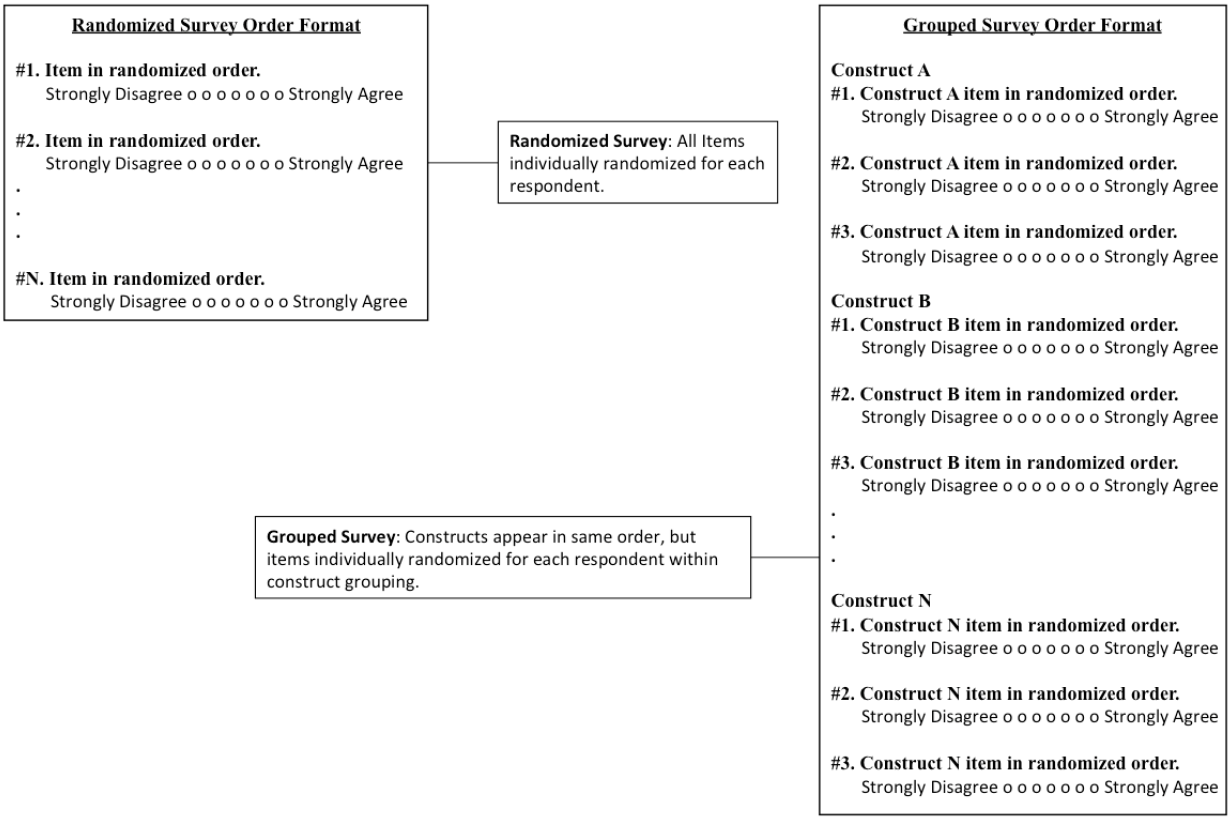

Figure 1. Survey Treatments

5 Results

We analyzed the data using SPSS 19 and AMOS 24. In examining the demographic data, we found an even split between female (50%) and male (50%) respondents. On average, respondents were 35 years old (youngest = 19; oldest = 84). A majority of participants (61.4%) had a bachelor's degree or higher.

5.1 Reliability

In order to test H1, we calculated Cronbach's alphas and composite reliabilities for each construct under each treatment (see Table 3). In all but one case (emotional appeal), the grouped treatment items had higher alphas compared to the randomized treatment items, and all significant differences between treatments resulted from higher values for grouped versus randomized reliability statistics. The overall average percentage difference between the random and grouped treatment was 5.86 percent (high of 14.9% for online completeness and low of -1.5% for emotional appeal). These results support H1 since we found inflated reliabilities in the grouped as compared to the randomized treatment for 92 percent (11 out of the 12) of the constructs. Using a non-parametric test, which assumes no systematic difference between alpha values, one would expect that the likelihood that a grouped treatment construct would have an alpha higher than the random treatment group to be 0.5 (50%). However, one would expect the likelihood that only one in 12 grouped treatments' reliabilities would exceed the random group to be 0.003 (based on the cumulative probability distribution). The fact that only emotional appeal had a higher reliability for random rather than grouped supports rejecting the null hypothesis that no systematic different between the treatments exists, which further supports H1.

Table 3. Reliability Analysis

Construct	Cronbach's alpha				Composite reliability		
	Random	Grouped	Significance		Random	Grouped	Significance
Informational fit-to-task	0.813	0.893	0.000		0.817	0.901	0.000
Tailored information	0.838	0.899	0.002		0.842	0.899	0.003
Trust	0.928	0.940	NS		0.930	0.941	NS
Response time	0.785	0.853	0.011		0.785	0.859	0.006
Ease of understanding	0.853	0.897	0.019		0.849	0.898	0.012
Intuitive operations	0.829	0.879	0.022		0.833	0.881	0.026
Visual appeal	0.958	0.978	0.000		0.958	0.979	0.000
Innovativeness	0.940	0.948	0.002		0.940	0.949	0.002
Emotional appeal	0.865	0.852	NS		0.887	0.879	NS
Consistent image	0.907	0.943	NS		0.909	0.943	NS
Online completeness	0.754	0.866	0.000		0.761	0.866	0.000
Relative advantage	0.756	0.846	0.003		0.769	0.879	0.000

5.2 Construct Validity

Next, in order to test for construct validity issues (H2), we conducted several analyses. First, we compared correlations to ensure grouped treatments did not contain artificially inflated reliabilities (Goodhue & Loiacono, 2002; Wilson & Lankton, 2012). For each construct in both treatments, we calculated an average of the items correlation with behavioral intentions along with its composite correlation with behavioral intentions. Consistent with previous research, we found that, on average, the calculated reliabilities were larger for the random treatment (0.05) and smaller for the grouped treatment (0.03) (see Table 4).

Table 4. Average of Individual Item Correlations, Composite Correlation

Construct	Random				Grouped		
	Avg. indv. item corr.	Composite corr.	Diff*		Avg. indv. item corr.	Composite corr.	Diff*
Informational fit-to-task	0.381	0.443	0.06		0.324	0.358	0.03
Tailored information	0.474	0.545	0.07		0.355	0.389	0.03
Trust	0.582	0.622	0.04		0.417	0.441	0.02
Response time	0.176	0.181	0.01		0.212	0.240	0.03
Ease of understanding	0.338	0.383	0.05		0.397	0.437	0.04
Intuitive operations	0.313	0.364	0.05		0.308	0.346	0.04
Visual appeal	0.642	0.669	0.03		0.589	0.601	0.01
Innovativeness	0.681	0.720	0.04		0.572	0.601	0.03
Emotional appeal	0.663	0.742	0.08		0.558	0.624	0.07
Consistent image	0.417	0.454	0.04		0.532	0.551	0.02
Online completeness	0.210	0.264	0.05		0.209	0.236	0.03
Relative advantage	0.242	0.294	0.05		0.276	0.314	0.04
Average			0.05				0.03

* Diff refers to the absolute difference between the average individual item correlation and composite correlation. Goodhue and Loiacono (2002) refer to this difference as "improvement due to uncorrelated errors canceling".

Second, we ran confirmatory factor analyses (CFA) for each sample (random = 200 participants and grouped = 202 participants) in order to identify whether they contained any discriminant and/or convergent

validity issues. We designed a measurement model in AMOS with each set of variables and their underlying construct identified.

Convergent validity issues exist when the constructs underlying variables (or items) do not correlate well with one another and show that the latent factor's observed variables do not explain it well. Conversely, if discriminant validity issues exist, then the variables correlate more highly with variables that measure another construct rather than their own parent factors (Hair et al., 2010). To determine the measurement models' convergent validity, we used the average variance extracted (AVE) measure. As Malhotra and Dash (2011) note:

> *Average variance extracted is a more conservative measure than composite reliability. On the basis of composite reliability alone, the researcher may conclude that the convergent validity of the construct is adequate, even though more than 50% of the variance is due to error.* (p. 702)

To evaluate discriminant validity, we compared the maximum shared variance (MSV) to the AVE and compared the square root of the AVE to the inter-construct correlations.

We depict the convergent and discriminant validity issues in each sample in Tables 5-7. We found the same number of construct validity issues (1) in both the random and grouped samples. However, the random sample had many more discriminant validity issues (14) compared to the grouped sample (2). Because the random and grouped data sets differed only in their item-ordering approach, these results indicate that grouped ordering artificially inflates construct validity statistics. Thus, we found support for H2.

5.3 Common Method Bias

We performed additional testing according to Podsakoff et al. (2003) to test for common method bias. Using AMOS, we included a common latent variable in the measurement model and added pathways to each variable. We constrained all paths to be the same. In the random treatment, the paths were 0.82, which resulted in common variance across the variables that equaled 67 percent. For the grouped treatment, the path was 0.94, which resulted in common variance across the variables that equaled 88 percent. We can attribute the higher level of common method variance for the grouped treatment to the survey format. Given that the measurement method differed only in that one treatment received items in random order and the other in construct groupings, we found support for H3. The grouped surveys had inflated common method bias effects compared to the randomized surveys.

Table 5. Construct Validity Measures: Random Sample

Construct	CR	AVE	MSV	MaxR(H)	OC	IFtT	TI	Tr	RT	EoU	IO	VA	Inn	EA	CI	RA
OC	0.761	0.516	0.773	0.767	**0.718**											
IFtT	0.817	0.598	0.918	0.888	0.879	**0.773**										
TI	0.842	0.640	0.912	0.931	0.767	0.955	**0.800**									
Tr	0.930	0.817	0.480	0.966	0.564	0.626	0.616	**0.904**								
RT	*0.465*	0.558	0.563	0.972	-0.674	-0.691	-0.614	-0.411	**0.747**							
EoU	0.829	0.618	0.918	0.976	0.750	0.958	0.791	0.472	-0.709	**0.786**						
IO	0.833	0.624	0.908	0.979	0.787	0.925	0.813	0.486	-0.750	0.953	**0.790**					
VA	0.958	0.885	0.696	0.987	0.381	0.677	0.696	0.546	-0.420	0.693	0.567	**0.941**				
Inn	0.940	0.839	0.696	0.989	0.299	0.527	0.609	0.513	-0.303	0.448	0.342	0.834	**0.916**			
EA	0.887	0.726	0.590	0.990	0.507	0.645	0.712	0.693	-0.386	0.512	0.521	0.768	0.755	**0.852**		
CI	0.909	0.770	0.555	0.991	0.636	0.745	0.693	0.578	-0.460	0.693	0.640	0.680	0.588	0.677	**0.877**	
RA	0.769	0.528	0.692	0.991	0.832	0.737	0.708	0.532	-0.583	0.684	0.801	0.414	0.268	0.479	0.585	**0.726**

Construct abbreviations: OC = online completeness; IFtT= informational fit-to-task; TI = tailored information; Tr = trust; RT = response time; EoU = ease of understanding; IO = intuitive operations; VA = visual appeal; Inn = innovativeness; EA = emotional appeal; CI = consistent image; RA = relative advantage.
Key: italics = convergent validity issue; gray = discriminant validity issue.

Table 6. Construct Validity Measures – Grouped Sample

Construct	CR	AVE	MSV	MaxR(H)	OC	IFtT	TI	Tr	RT	EoU	IO	VA	Inn	EA	CI	RA
OC	0.866	0.683	0.381	0.869	**0.826**											
IFtT	0.901	0.753	0.752	0.945	0.575	**0.868**										
TI	0.899	0.748	0.701	0.963	0.464	0.837	**0.865**									
Tr	0.941	0.843	0.332	0.978	0.368	0.480	0469	**0.918**								
RT	*0.471*	0.670	0.371	0.981	-0.454	-0.542	-0.542	-0.265	**0.819**							
EoU	0.898	0.746	0.638	0.984	0.351	0.678	0.634	0.484	-0.401	**0.864**						
IO	0.881	0.712	0.752	0.985	0.559	0.867	0.766	0.442	-0.609	0.799	**0.844**					
VA	0.979	0.938	0.529	0.992	0.249	0.344	0.434	0.438	-0.278	0.661	0.399	**0.969**				
Inn	0.949	0.861	0.529	0.993	0.236	0.352	0.441	0.525	-0.248	0.527	0.360	0.727	**0.928**			
EA	0.879	0.714	0.399	0.994	0.236	0.352	0.441	0.505	-0.290	0.465	0.351	0.615	0.616	**0.845**		
CI	0.943	0.847	0.399	0.994	0.516	0.419	0.447	0.576	-0.248	0.542	0.471	0.616	0.583	0.632	**0.920**	
RA	0.879	0.714	0.381	0.995	0.617	0.530	0.492	0.295	-0.362	0.437	0.601	0.236	0.107	0.268	0.412	**0.845**

Construct abbreviations: OC = online completeness; IFtT= informational fit-to-task; TI = tailored information; Tr = trust; RT = response time; EoU = ease of understanding; IO = intuitive operations; VA = visual appeal; Inn = innovativeness; EA = emotional appeal; CI = consistent Image; RA = relative advantage.
Key: italics = convergent validity issue; gray = discriminant validity issue.

Table 7. Discriminant Validity Issues

Construct	Random sample		Grouped sample	
	Issue 1	Issue 2	Issue 1	Issue 2
Informational fit-to-task	X	X		
Tailored information	X	X		
Trust				
Response time	X	X		
Ease of understanding	X	X		
Intuitive operations	X	X	X	X
Visual appeal				
Innovativeness				
Emotional appeal				
Consistent image				
Online completeness	X	X		
Relative advantage	X	X		

Issue 1: the square root of the AVE for the identified construct is less than at least one of the absolute values of the correlation with another construct
issue 2: the AVE for the identified construct is less than the MSV.

5.4 Semantic Effects

As we mention in Section 3, logically grouping items can facilitate respondents' cognitive processing by making meanings clearer or memory retrieval easier. However, grouping items in this way can introduce semantic effects. In order to measure the impact of semantic effects in this study, we compared the item mean scores for the random and group treatments using a one-way ANOVA test. Of the 36 item mean comparisons, we found 12 mean differences and nine statistically significant comparisons. In 11 mean differences, the grouped mean was higher than the random mean (see Appendix A).

Again, using a non-parametric test, we examined the null hypothesis that asserted that 50 percent of the means would be higher for the random sample and for the grouped treatment. The cumulative probability of 12 or fewer successes in 36 attempts was 0.032, which rejects the null. We found a significant difference between the two groups' item means and, thus, support for H4.

5.5 Sequential Effects

In order to test for sequential effects (H5), we ran several analyses. First, we conducted a correlation test between the average responses in the grouped condition and construct order (i.e., 1-12) to assess whether responses became biased in a particular direction during the survey. We did not appear to find a significant bias trend ($p = 0.097$). Second, we conducted another correlation test between average difference scores (grouped - random) and construct order to assess whether the direction of differences changed during the survey. We found the value of grouped - random differences became more negative as the survey progressed ($p < 0.000$). Similarly, a correlation test between absolute grouped - random absolute difference scores showed that the magnitude of differences between conditions reduced as the survey progressed ($p = 0.049$). Thus, we found support for H5. We report specific items and their means and differences in Table 8.

Table 8. Item Means and Differences

Item	Grouped	RG = 0	RG = 1	Difference	ABS difference
	Construct order	Random	grouped	Group - rand	ASB (grouped - random)
INFO_1	1	5.91	6.25	0.34	0.34
INFO_2	1	5.72	6.13	0.41	0.41
INFO_3	1	5.86	6.05	0.19	0.19
TAILOR_1	2	5.55	5.92	0.37	0.37
TAILOR_2	2	5.45	5.89	0.44	0.44
TAILOR_3	2	5.69	5.93	0.24	0.24
TRUST_1	3	4.98	5.07	0.09	0.09
TRUST_2	3	4.79	4.98	0.19	0.19
TRUST_3	3	4.82	4.95	0.13	0.13
RESP_1	4	5.88	5.95	0.07	0.07
RESP_2	4	5.95	6.01	0.06	0.06
RESP_3	4	5.82	6.03	0.21	0.21
EUDSTD_1	5	5.75	5.81	0.06	0.06
EUDSTD_2	5	5.79	5.93	0.14	0.14
EUDSTD_3	5	5.88	5.91	0.03	0.03
INTUIT_1	6	6.10	6.20	0.10	0.10
INTUIT_2	6	5.97	6.24	0.27	0.27
INTUIT_3	6	5.96	6.14	0.18	0.18
VISUAL_1	7	4.77	4.79	0.02	0.02
VISUAL_2	7	4.75	4.80	0.05	0.05
VISUAL_3	7	4.79	4.78	-0.01	0.01
INNOV_1	8	4.25	4.23	-0.02	0.02
INNOV_2	8	4.20	4.10	-0.10	0.10
INNOV_3	8	4.53	4.32	-0.21	0.21
EMOTION_1	9	4.61	4.47	-0.14	0.14
EMOTION_2	9	4.52	4.44	-0.08	0.08
EMOTION_3	9	3.73	3.65	-0.08	0.08
CONSIMG_1	10	5.33	5.07	-0.26	0.26
CONSIMG_2	10	5.23	5.03	-0.20	0.20
CONSIMG_3	10	5.33	5.07	-0.26	0.26
OLCOMP_1	11	6.18	6.02	-0.16	0.16
OLCOMP_2	11	5.76	5.79	0.03	0.03

Table 8. Item Means and Differences

OLCOMP_3	11	5.75	5.78	0.03	0.03
RELADV_1	12	5.93	6.19	0.26	0.26
RELADV_2	12	6.09	6.15	0.06	0.06
RELADV_3	12	5.64	5.73	0.09	0.09

5.6 Survey Effects on Participant

In order to test H6, we compared frustration and fatigue measures (Bradly & Daly, 1994; Caussade et al., 2005; Hess et al., 2012; Smith, Smith, Gray, & Ryan, 2007) for both treatment groups. We used three items to measure participants' fatigue and three to determine their frustration. We presented these items to participants at the end of the survey (the scale alpha reliability of the three fatigue items was 0.966 and 0.766 for frustration). We calculated the summated mean for each multi-measure and used it to perform a means comparison test in SPSS. Though the means for the random treatment fatigue (1.97) and frustration (2.66) were slightly higher compared to the group treatment (1.78 and 2.55, respectively), we found no significant difference between them (see Tables 9 and 10). Thus, we did not find support for H6.

Table 9. Items Measuring Participants' Mental State Following Trust Study*

Item	Text	Scale alpha
Fatigue 1	Completing the questionnaire has made me feel worn out.	0.966
Fatigue 2	Completing the questionnaire has made me feel fatigued.	
Fatigue 3	Completing the questionnaire has made me feel weary.	
Frustration 1	Being asked similar questions over and over in the questionnaire frustrates me.	0.766
Frustration 2	The number of questions I am being asked in the questionnaire makes me feel frustrated.	
Frustration 3	After completing the questionnaire, I am not frustrated at all. *(Reverse coded)*	

* We individually randomized the order we administered these questions to participants in both conditions. We measured all items using a seven-point Likert scale (1 = strongly disagree, 2 = somewhat disagree, 3 = disagree; 4 = neutral; 5 = disagree, 6 = somewhat agree, 7 = strongly agree).

Table 10. Participant Fatigue and Frustration

Construct	Mean (st. dev.)		Test		
	Random	Grouped	F-test	Sig.	Partial Eta^2
Fatigue	1.97 (1.22)	1.78 (0.99)	3.11	0.078	0.008
Frustration	2.66 (1.46)	2.55 (1.37)	0.616	0.433	0.002

In addition to collecting frustration and fatigue measures, we included an open-ended question at the end of the survey that asked respondents: "Is there anything else you would like to say about the survey you just took?" Overall, respondents in both treatment groups made similar comments. We highlight common respondent comments to the open-ended question in Table 11.

Table 21. Common Participant Comments

Response	Survey type	
	Random	Grouped
Left blank	72	80
"No", "Nope", "Thank you", or "Nothing I can think of"	75	70
Positive comment	14	13
Negative comment	3	4

Participants in both random and grouped treatments clearly noted repetitive questions as a negative, but they acknowledged that it was not overtly frustrating. For example, in the random treatment, a respondent commented: "Although the survey was a bit repetitive it was short enough as to not become boring". In the grouped treatment, a respondent noted: "There were quite a lot of repetitive questions, but that seems fair because sometimes wording can change opinions". Ironically, questions that asked if the respondents found the survey frustrating or fatiguing seemed to annoy them more. A random treatment member stated" "I liked it besides the constant questions of if the survey itself was annoying or not". Similarly, a grouped treatment member said: "The survey did not annoy me, but asking me [at the end of the survey] if I was annoyed over and over again did". We provide specific participant feedback on the randomized versus grouped survey formats in Tables 12 and 13. We highlight the results of all hypotheses in Table 14.

Table 32. Positive Participant Comments

Random	Grouped
• It was fun and thought provoking. • Everything was pretty clear with the survey. • Very enjoyable—thanks for the opportunity! • It was interesting… • Was much faster and easier than others I have taken. Repeating questions is ok if you don't overdo it. Thanks for the work • Very easy to understand and complete. Thank you. • Actually it wasn't as bad as a lot I have taken • Although the survey was a bit repetitive it was short enough as to not become boring… • …Questions at the end were frustrating, not the survey • I always enjoy tricky surveys :) • Well done and enjoyable...	• Interesting survey • Interesting survey… • Simple and quick • Interesting and enjoyable • It was really easy and more interesting than most. • This was a great survey… • It was interesting. • The survey was quick and the instructions were very clear. • …It was fun… • Nice to have a survey that is a little bit different. • Interesting survey, I enjoyed taking it, thanks! • Interesting survey with clear easy to follow directions… • Interesting survey.

Table 43. Negative Participant Comments

Random	Grouped
• I'm a nitpicky person, so a slight change in wording of a statement made me consider how it was different from those that sounded similar, and whether the change in wording changed my answer. • I liked it besides the constant questions of if the survey itself was annoying or not. • Does this question annoy you? If that one doesn't, does this one cause a slight nuisance? Would it be rather aggravating to have to answer this question too? Is this question the most frustrating of them all? …Are you trying to convince yourself that this actually is productive? Are you hypnotizing yourself into believing that pointless tasks are productive? Are you ready to answer a few more questions?	• There were quite a lot of repetitive questions, but that seems fair because sometimes wording can change opinions • These are basically the same questions re-worded three times. • It seemed that the questions about whether I was annoyed by the survey wanted me to say yes. In other words, the survey did not annoy me, but asking me if I was annoyed over and over again did. • It was a bit redundant…

Table 54. Hypotheses Results

#	Hypothesis	Supported?
H1	Grouped surveys have inflated reliability compared to randomized surveys.	Yes
H2	Grouped surveys have inflated construct validity statistics compared to randomized surveys.	Yes
H3	Grouped surveys have increased common method bias effects compared to randomized surveys.	Yes
H4	Semantic context effects occur in in the grouped survey format.	Yes
H5	Sequential context effects occur in the grouped survey format.	Yes
H6	Grouped survey participants have lower fatigue and frustration compared to randomized survey participants.	No

6 Discussion

6.1 Research Implications

Researchers have debated randomization's impact for decades but conducted little research to help resolve the issues related to the true impact that non-item randomization has on research results and participants. From our study, we can see the significant impact that fully randomized surveys can have in maintaining IS research's veracity and integrity. Even semi-randomization, as represented in our grouped survey treatment, did not negate the effects of inflated reliability, validity, and common method bias. These findings not only confirm past studies (Goodhue & Loiacono, 2002; Wilson & Lankton, 2012; Wilson et al., 2017) but also extend them to more fully explain the substantial impact grouping can have on survey results. Even with intra-correlation item randomization, fixed correlation placement can have a significant effect on study results.

Additionally, by employing adults who ranged from 18 to 81 years old, we extend previous research by increasing the extent to which our results generalize to the larger adult population. Though U.S. undergraduate student participants represent an acceptable participant pool for such research, it limits the research since such a group is much younger than the general adult population and may differ in their feelings about question organization.

Further, the similarity in respondent comments between the two treatment groups adds to our understanding about what additional effort randomization may put on respondents. Fatigue and frustration between the two groups did not significantly differ, and the comments affirm that questions that asked if the respondents found the survey frustrating or fatiguing seemed to annoy them more than "repetitive" questions. In fact, the grouped questions seemed to make them more aware about the repetitive questions than if we had randomly placed similar items throughout the survey. Researchers with concerns that randomization may negatively impact their results may find these results especially helpful and illuminating.

6.2 Practical Implications

From a practical perspective, surveys that researchers develop under complete item randomization have more accurate measurements. The randomization negates the method bias that could affect measures. Thus, researchers and practitioners employing these measures can feel more confident in their results.

Having such accurate measures also represents an essential part of ensuring replicability, a central tenant in the scientific method. Recently, however, a crisis has arisen in the social sciences related to replication (Pashler & Wagenmakers, 2012). Researchers in many fields, including IS, have found it difficult to replicate previous studies. Often, the replicated studies end with contradictory results, which can pose problems for a field, such as HCI, when researchers base fundamental theories on experimental work. As Dennis and Valacich (2014) note, "replications are valuable in advancing science". In order to conduct true replications, researchers and practitioners must be clear about the measurements they use.

6.3 Limitations and Future Research

With this research, we expand our knowledge about grouping effects on survey items. However, additional work can still explore the effect that different semantic-differential Likert measures have on statistical inflation when using grouped versus randomized survey items. Where this research and previous research (e.g., Wilson & Lankton, 2012) used a seven-position semantic-differential, a five-point Likert measure may reveal different results.

Additionally, we understand that researchers may sometimes lack the freedom to completely randomize their survey items and so need to group items. Such instances may include the need to maintain methodological integrity with prior administrations or to rigorously apply a grouped-static validated instrument. In these cases, researchers need to detail their approach for fellow researchers, who may need to replicate their research method.

7 Conclusion

Our research results strongly suggest that researchers should use fully randomized survey instruments in HCI research. Grouped ordering of any kind increases bias and statistical inflation, which can influence results' veracity. However, researchers can more easily control these effects with full randomization. Additionally, we did not find any significant increase in frustration or fatigue in participants in both the randomized and grouped treatments. Thus, completely randomized surveys should be the de facto survey that IS research uses, especially given the proliferation of online surveying options. Editors and reviewers in this regard should encourage transparency as well: they should encourage authors to be up front in their method choice and explain why they may have varied from a purely randomized approach.

References

Arnetz, B. B., & Berg, M. (1993). Techno-Stress. psycho-physiological consequences of poor man-machine interface. In M. J. S. G. Salvendy (Ed.), *Human-computer interaction: Applications and case studies* (pp. 891-896). Amsterdam: Elsevier.

Backor, K., Golde, S., & Nie, N. (2007). *Estimating survey fatigue in time use study.* Paper presented at the 2007 International Association for Time Use Research Conference, Washington, DC.

Bradlow, E. T., & Fitzsimons, G. J. (2001). Subscale distance and item clustering effects in self-administered surveys: A new metric. *Journal of Marketing Research*, *38*(2), 234-261.

Bradley, M., & Daly, A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation*, *21*(2), 167-184.

Brod, C. (1984). *Technostress: The human cost of the computer revolution.* Reading, MA: Addison-Wesley.

Buchanan, E. M., Foreman, R. E., Johnson, B. N., Pavlacic, J. M., Swadley, R. L., & Schulenberg, S. E. (2018). Does the delivery matter? Examining randomization at the item level. *Behaviormetrika, 45*(2), 1-22.

Budd, R. J. (1987). Response bias and the theory of reasoned action. *Social Cognition*, *5*, 95-107.

Campbell, D. T. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, *56*(2), 81-105.

Caussade, S., Ortúzar, J. D., Rizzi, L. I., & Hensher, D. A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research Part B*, *39*(7), 621-640.

Chang, S., van Witteloostuijn, A., & Eden, L. (2010). From the editors: Common method variance in international business research. *Journal of International Business Studies*, *41*(2), 178-184.

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly*, *65*(2), 230-253.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340.

Davis, F. D., & Venkatesh, V. (1996). A critical assessment of potential measurement biases in the technology acceptance model: Three experiments. *International Journal of Human-Computer Studies*, *45*(1), 19-45.

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

Goodhue, D. L., & Loiacono, E. T. (2002). Randomizing survey question order vs. grouping questions by construct: An empirical test of the impact on apparent reliabilities and links to related constructs. In *Proceedings of the 35th Hawaii International Conference on System Sciences.*

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 155-167.

Hair, J., Black, W., Babin, B., & Anderson, R. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice-Hall.

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. *Journal of Applied Psychology*, *78*(1), 129-140.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549-559.

Hess, T. J., McNab, A. L., & Basoglu, K. A. (2014). Reliability generalization of perceived ease of use, perceived usefulness, and behavioral intentions. *MIS Quarterly*, *38*(1), 1-28.

Johnson, W. R., Sieveking, N. A., & Clanton, E. S. (1974). Effects of alternative positioning of open-ended questions in multiple-choice questionnaires. *Journal of Applied Psychology*, *59*(6), 776-778.

Kane, M. T. (2008). *Errors of measurement, theory, and public policy.* Paper presented at the 12th Annual William H. Angoff Memorial Lecture New Jersey.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*(2), 312-320.

Knowles, E. E., & Byers, B. (1996). Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology*, *70*(5), 1080-1090.

Kraut, A. I., Wolfson, A. D., & Rothenberg, A. (1975). Some effects of position on opinion survey items. *Journal of Applied Psychology*, *60*(6), 774-776.

Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed.). West Yorkshire, England: Emerald Group.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods.* Thousand Oaks, CA: Sage.

Lazarus, R. S. (1966). *Psychological stress and the coping process.* New York, NY: McGraw-Hill.

Lazarus, R. S. (1993). Why we should think of stress as a subset of emotion. In L. G. S. Breznitz (Ed.), *Handbook of stress: Theoretical and clinical aspects* (pp. 21-39). New York, NY: Free Press.

Loiacono, E. T., Watson, R. T., & Goodhue, D. L. (2007). WebQual: An instrument for consumer evaluation of web sites. *International Journal of Electronic Commerce*, *11*(3), 51-87.

Loiacono, E. T. & Watson, R., & Goodhue, D. L. (2002). WEBQUAL: A measure of website quality. In *Proceedings of the Marketing Educators' Conference: Marketing Theory and Applications* (pp. 432-437).

Lowry, P. B., Gaskin, J. E., & Moody, G. D. (2015). Proposing the multimotive information systems continuance model (MISC) to better explain end-user system evaluations and continuance intentions. *Journal of the Association for Information Systems*, *16*(7), 515-579.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, *35*(2), 293-334.

Malhotra, N. K., & Dash, S. (Ed.). (2011). *Marketing research: An applied orientation.* New Delhi: Pearson.

Martin, E. (1980). The effects of item contiguity and probing on measures of anomia. *Social Psychology Quarterly*, *43*(1), 116-120.

McCroskey, J. C., Richmond, V. P., & McCroskey, L. L. (2006). *An introduction to communication in the classroom: The role of communication in teaching and training.* Boston, MA: Pearson.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528-530.

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

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012) Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539-569.

Raghunathan, T. E., & Grizzle, J. E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, *90*(429), 54-63.

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Schriesheim, C. A., Solomon, E., & Kopelman, R. E. (1989). Grouped versus randomized format: An investigation of scale convergent and discriminant validity using LISREL confirmatory factor analysis. *Applied Psychological Measurement*, *13*(1), 19-31.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, working, and content.* New York, NY: Academic Press.

Schwarz, N., & Sudman, S. (2012). *Context effects in social and psychological research.* New York, NY: Springer-Verlag.

Sijtsma, K. (2009). Over misverstanden rond Cronbachs alfa en de wenselijkheid van alternatieven. *Psycholoog*, *44*(11), 561-567.

Smith, T. W. (1983). The hidden 25 percent: An analysis of nonresponse in the 1980 General Social Survey. *Public Opinion Quarterly*, *47*(3), 386-404.

Smith, T. W. (1988). Context effects in the general social survey. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds), *Measurement errors in surveys* (pp. 57-72). New York, NY: Wiley.

Smith, B., Smith, T. C., Gray, G. C., & Ryan, M. A. (2007). When epidemiology meets the Internet: Web-based surveys in the millennium cohort study. *American Journal of Epidemiology*, *166*(11), 1345-1354.

Straub, D., Boudreau, M.-C., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems*, *13*, 380-427.

SurveyGizmo (2015). *Logic, piping, repeating and randomization.* Retrieved from https://help.surveygizmo.com/help/logic-piping

Taylor, J. (1999). *An introduction to error analysis: The study of uncertainties in physical measurements.* Sausalito, CA: University Science Books.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*(3), 299-314.

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

Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*(3), 368-393.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185-201.

Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing*, *26*(1), 2-12.

Weinberger, A. H., & Darkes, J., Del Boca, F. K., Greenbaum, P. E., & Goldman, M. S. (2006). Items as context: Effects of item order and ambiguity on factor structure. *Basic and Applied Social Psychology*, *28*(1), 17-26.

Williams, L. J., & Brown, B. K. (1994). Method variance in organizational behavior and human resources research: Effects on correlations, path coefficients, and hypothesis testing. *Organizational Behavior and Human Decision Processes*, *57*, 185-209.

Wilson, E. V., & Lankton, N. K. (2012). Some unfortunate consequences of non-randomized, grouped-item survey administration in IS research. In *Proceedings of the 2012 International Conference on Information Systems.*

Wilson, E. V. Srite, M., & Loiacono, E. T. (2017). A call for item-ordering transparency in online IS survey administration. In *Proceedings of the 2017 Americas Conference on Information Systems.*

Appendix A: Instrumentation Detail

Table A1. Item Mean Differences between Randomized and Grouped Treatments

Construct	#	Item	Means		Test	
			Random	Group	F	Sig
Informational fit-to-task	1	The information on the website is pretty much what I need to carry out my tasks.	5.91	6.25	12.864	0.000
	2	The website adequately meets my information needs.	5.72	6.13	14.707	0.000
	3	The information on the website is effective.	5.86	6.05	3.777	0.053
Tailored information	4	The website allows me to interact with it to receive tailored information.	5.55	5.92	10.193	0.002
	5	The website has interactive features, which help me accomplish my task.	5.45	5.89	14.493	0.000
	6	I can interact with the website in order to get information tailored to my specific needs.	5.69	5.93	4.971	0.026
Trust	7	I feel safe in my transactions with the website.	4.98	5.07	0.503	0.479
	8	I trust the website to keep my personal information safe.	4.79	4.98	2.065	0.151
	9	I trust the website administrators will not misuse my personal information.	4.82	4.95	0.89	0.346
Response time	10	When I use the website there is very little waiting time between my actions and the website's response.	5.88	5.95	0.304	0.582
	11	The website loads quickly.	5.96	6.01	0.238	0.626
	12	The website takes long to load.	2.18	1.98	2.484	0.116
Ease of Understanding	13	The display pages in the website are easy to read.	5.75	5.81	0.369	0.544
	14	The text on the website is easy to read.	5.79	5.93	1.677	0.196
	15	The website labels are easy to understand.	5.89	5.91	0.035	0.851
Intuitive operations	16	Learning to operate the website is easy for me.	6.11	6.20	0.938	0.333
	17	It would be easy for me to become skillful at using the website.	5.97	6.24	7.227	0.007
	18	I find the website easy to use.	5.96	6.14	2.662	0.104
Visual appeal	19	The website is visually pleasing.	4.77	4.79	0.018	0.894
	20	The website displays visually pleasing design.	4.75	4.80	0.083	0.774
	21	The website is visually appealing.	4.79	4.78	0.002	0.963
Innovative-ness	22	The website is innovative.	4.25	4.23	0.01	0.918
	23	The website design is innovative.	4.20	4.10	0.357	0.550
	24	The website is creative.	4.53	4.32	1.66	0.198
Emotional appeal	25	I feel happy when I use the website.	4.61	4.47	1.119	0.291
	26	I feel cheerful when I use the website.	4.53	4.44	0.432	0.512
	27	I feel sociable when I use the website.	3.73	3.65	0.227	0.634
Consistent image	28	The website projects an image consistent with the company's image.	5.33	5.07	4.728	0.030
	29	The website fits with my image of the company.	5.24	5.03	2.953	0.087
	30	The website's image matches that of the company.	5.33	5.07	4.702	0.031

Table A1. Item Mean Differences between Randomized and Grouped Treatments

Online completeness	31	The website allows transactions online.	6.18	6.02	1.991	0.159
	32	All my business with the company can be completed via the website.	5.76	5.79	0.073	0.786
	33	Most all business processes can be completed via the website.	5.76	5.78	.035	0.851
Relative advantage	34	It is easier to use the website to complete my business with the company than it is to telephone, fax, or mail a representative.	5.93	6.19	5.345	0.021
	35	The website is easier to use then calling an organizational representative agent on the phone.	6.09	6.15	.305	0.581
	36	The website is an alternative to calling customer service or sales.	5.64	5.73	.537	0.464
Note: we measured all items using a seven-point Likert scale where 1 = “strongly disagree” and 7 = “strongly agree”.						

About the Authors

Eleanor T. Loiacono is Professor of Information Technology and Data Science and the Co-Founder and Director of the Inclusive Design and Accessibility (IDEA) Hub (idea.wpi.edu) at Worcester Polytechnic Institute. Her expertise centers on the intersection of technology and the user. Over the past 20 years, she has focused on how people feel about the technology they use and how technologies, such as mobile apps and social media, can improve users' experiences. She is particularly interested in how those with differing abilities interact with technologies.

E. Vance Wilson is an Associate Teaching Professor in the Foisie School of Business at Worcester Polytechnic Institute. His research focuses on organizational aspects of human-computer interaction with special emphasis on e-health, computer-mediated communication, and persuasion.

Transactions on Human – Computer Interaction