TPM®

IMPACT OF QUESTIONNAIRE FORMAT ON RELIABILITY, VALIDITY, AND HYPOTHESIS TESTING

MADHUBALAN VISWANATHAN
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

UJWAL KAYANDE
MELBOURNE BUSINESS SCHOOL

RICHARD P. BAGOZZI
UNIVERSITY OF MICHIGAN

SAM RIETHMULLER
COLES FLYBUYS, MELBOURNE

SHIRLEY Y. Y. CHEUNG
HONG KONG BAPTIST UNIVERSITY

Researchers use a variety of questionnaire formats to collect data on measures of constructs for theory testing. For example, researchers may label measures of constructs, present measures of different constructs on different pages, or intersperse items from different constructs. Such questionnaire format choices are often guided by commonly held beliefs and conventions. However, there is little if any empirical research evaluating such different formats, precluding an informed view of the appropriateness of a format for a study. To our knowledge, our research represents the first systematic empirical investigation of how different formats affect research outcomes. We conduct a series of studies to systematically examine the effects of questionnaire format on the psychometric properties of measures of constructs and the relationships between constructs. Using multiple group confirmatory factor analysis (MGCFA), we find that the measures are largely invariant to questionnaire format when using student samples, and recommend that researchers should reconsider the rationale provided for choosing specific formats.

Key words: Measurement; Reliability; Theory-testing; Multiple group confirmatory factor analysis; Questionnaire formats.

Correspondence concerning this article should be addressed to Madhubalan Viswanathan, University of Illinois Urbana-Champaign, 183 Wohlers Hall, 1206 South Sixth Street, Champaign, IL 61820, USA. Email: mviswana@illinois.edu

Researchers commonly use questionnaires with individual or organizational foci and collect data on a variety of measures of constructs. Such questionnaires can be structured in many different ways. Measures of dimensions of a construct could be labeled, or presented separately on different pages. Items from measures of multiple dimensions or even different constructs could be interspersed. The sequencing of items within a measure of a dimension could be changed. Although

the actual layout of measures in a questionnaire is not commonly discussed in a published paper, researchers tend to make a conscious decision about the structure — that is, format — based on commonly held beliefs and conventions.

Underlying the different practices followed by researchers is the rationale that each of these practices ameliorates a measurement-related problem that eventually affects research outcomes (e.g., relationships between variables in tests of hypotheses). Yet, there has not been a systematic empirical investigation of whether and how different format practices affect research outcomes. As a consequence, systematic guidance on the choice of questionnaire format is very limited. In essence, the lack of systematic research in this area calls into question the reliability and validity of measures and inferences about relationships between those constructs in past research, based on the varying use of practices such as labeling or interspersing. Given the lack of empirical work on this topic, current research design practice rests on conceptual arguments and implicit knowledge about possible outcomes of different questionnaire format alternatives.

The task of identifying methodological factors that affect reliability and validity is itself challenging, and relevant factors may often be considered too idiosyncratic to study systematically. Moreover, much of current knowledge or working hypotheses in this realm may well be tacit or implicit in nature, based on individual researchers' experiences. However, such tacit working hypotheses need to be explicated and studied systematically. Our main objective here is to systematically examine the effects of methodological factors stemming from questionnaire format on the psychometric properties of measures of constructs and on the relationships between constructs. Specifically, we investigate the impact of (i) sequencing of items within and across measures, and (ii) grouping or separation of items and measures through labeling, pagination, or contiguous placement. Additionally, we investigate whether questionnaire length (i.e., the burden placed on respondents), different types of response scales (respondent- or stimulus-centered), or different types of samples (student versus non-student adults) moderate the impact of format on psychometric properties of measures of constructs.

In placing this research in perspective, research on the effects of a variety of factors on responses in survey research with its focus on estimating accurate means (e.g., Sudman, Bradburn, & Schwarz, 1996) provides a noteworthy parallel. This literature has systematically identified and studied a number of factors that influence responses to individual questions. In this paradigm, the emphasis has been on eliciting unbiased responses that do not deviate from the "true" value. However, such "additive error" is relatively less problematic in research, such as in many types of academic research, which emphasizes accurate estimates of *relationships* between variables rather than accurate estimates of absolute values per se (Groves, 1991). Relevant for such research is an understanding of methodological factors relating to format, such as labeling of constructs, on the *psychometric properties of measures, as well as on the relationships between constructs*. Relationships between constructs are typically not affected by some additive error, rather, a nuanced understanding of correlational systematic error is required (Viswanathan, 2005) — error that affects associations between variables.

Our research is in the tradition of methodological articles on such topics as common method variance (Cote & Buckley, 1988; Malhotra, Kim, & Patil, 2006; Podsakoff, Podsakoff, MacKenzie, & Lee, 2003) and the effects of measure design (e.g., measure development process), sample characteristics, and scale design on the reliability of rating scales (Churchill & Peter, 1984;

Weng, 2004). We differ, however, in our focus on how format factors, such as labeling of constructs or interspersion of items from different constructs, affect reliability and validity. Thus, our work lies at the middle of the continuum from examining common method variance at one end (i.e., at the unit of analysis of the overall method) and examining micro-level measure design or scale properties, such as response category anchors or number of response categories at the other end (i.e., at the level of response scales or scale development procedures as the unit of analysis). There is, of course, some overlap, between the format factors we study and research on either end of this continuum — for example, intermixing of items of constructs in the context of common method variance (Kline, 2000; Podsakoff et al., 2003) referred to here as interspersion, and time and location of measurement of constructs (Podsakoff et al., 2003) also addressed here. We also note that Bradlow and Fitzsimons (2001) represent one exception in having examined what we refer to as format factors of labeling and grouping, discussed subsequently.

This paper is organized as follows. Following a general articulation of measurement error, we discuss different methodological practices in terms of their assumed effect on types of measurement error. We report on a number of empirical studies. We introduce an application of multiple group confirmatory factor analysis to test whether these assumptions are borne out empirically, or whether measures are invariant across the different methodological practices. We then discuss five studies to test these assumptions, followed by a description of the analysis and results. Finally, we conclude with specific prescriptions for empirical research.

Measurement Error and Methodological Factors in Research

The implicit rationale used by researchers in choosing one format over another is that it ameliorates a potential measurement-related problem arising from measurement error. Here, we discuss different types of measurement error that could occur and how each might affect the observed relationship between variables. We then discuss methodological practices relating to format and discuss how such practices are assumed to impact measurement. We conclude this section by introducing multiple group confirmatory factor analysis as an approach to empirically test whether commonly used practices actually impact the psychometric properties of measures.

Measurement Error

We first present a review of the different types of measurement error by way of background for the rest of the paper. *Additive systematic error* occurs with deviations from the true score by a constant magnitude (e.g., extreme means) and may influence observed relationships when decreased item variance reduces covariation with other items (Viswanathan, 2005). For research with a focus on relationships between constructs, this type of systematic error may be less problematic when compared to contexts, such as opinion research, where the onus is on estimating absolute values (Groves, 1991).

With *correlational systematic error*, responses vary consistently and by different degrees beyond true differences in the measured construct (Viswanathan, 2005). Response categories, such as *very good* to *very bad*, may be used in consistent but different ways by different individuals,

wherein *very bad* is more or less negative for different respondents. Correlational systematic error may strengthen or weaken observed relationships (Nunnally, 1978).

Within-measure correlational systematic error arises between different items of the same construct, with examples including stronger observed relationships between items of a construct when using the same response format (Viswanathan, 2005). Halo error is an example wherein a global impression is employed to complete ratings on measures of distinct dimensions (Lance, LaPointe, & Stewart, 1994). A "halo" can be created by responses to one or two items, and lead to consistent responses to other items of the construct. Responses to later items in a scale may, therefore, be more polarized and consistent (Feldman & Lynch, 1988; Knowles, 1988; Simmons, Bickart, & Lynch, 1993) and more reliable, by following responses to earlier items (Knowles & Byers, 1996).

Across-measure correlational systematic error leads to inaccurate but consistent observed relationships, increasing or decreasing correlations (Viswanathan, 2005). Common method factors, such as placement of items of different measures on one page (Lennox & Dennis, 1994), represent examples of this type of error.

Methodological Practices and Measurement Error

Multiple practices are typically followed by researchers in the social sciences, in structuring their questionnaires. Items measuring different constructs can be interspersed — for example, Parameswaran, Barnett, Greenberg, Bellenger, and Robertson (1979) interspersed items from different domains of lifestyles; Szybillo, Binstok, and Buchanan (1979) interspersed items measuring attitude importance with other items; please also see Smith, Haugtvedt, and Petty (1994). Items measuring a construct can be placed contiguously with or without labels (e.g., Bradlow & Fitzsimons, 2001), or items measuring different constructs can be placed on different pages (e.g., Mittal 1995). Each of these practices may have an impact on correlational systematic error, leading to a potential effect on the estimates of relationships between constructs. For example, items that are used to test theories of relationships between constructs can be interspersed to mitigate inflation of correlations that could result from halo error. However, as these practices have received little systematic inquiry a clear theoretical rationale is not available for why researchers use one format over another. Thus, rather than express formal hypotheses about the effects of each condition, we present typical arguments or assumptions currently held in practice and in the literature.

We examined a total of eight conditions that varied on two methodological factors: (i) item sequencing within and across constructs, and (ii) separation or grouping of items and constructs through labeling, pagination, or contiguous placement. Treating the contiguous placement of items and measures (for multidimensional constructs) as the baseline condition, we then compared the effect of adding labels to measures, paginating measures, and interspersing (or resequencing) items and measures. We present a summary of the conditions in the Appendix A.

In Condition 1, referred to as "contiguous," items from each construct/dimension (the latter if the construct is multidimensional) are presented in proximity contiguously, a common practice and a useful baseline. Condition 2, referred to as "contiguous and labeled," is similar to Condition 1 with the addition of labels for measures/subscales of unidimensional constructs/dimensions of multidimensional constructs. In Condition 3, referred to as "contiguous and paginated," measures/subscales for each construct/dimension are presented on a different page, without labels. Thus,

Conditions 2 and 3 test the effects of labeling and pagination, respectively. Each of these approaches serves to provide a logical division between measures/subscales of dimensions of a construct and the constructs themselves, labeling more explicitly and pagination more subtly. Such division may lead to greater consistency among items within measures of specific constructs, or within subscales of specific dimensions of a construct, exploiting within-measure correlational systematic error. Stability reliability may also be enhanced, due to consistency over time among items. Because subscales of individual dimensions, rather than the measure of the overall multidimensional construct, are labeled, items are expected to have higher loadings on respective factors representing dimensions. Labeling may have stronger effects than pagination due to the explicit naming of constructs and dimensions. In this regard, Bradlow and Fitzsimons (2001) found that labeling and grouping (of items on a screen), each led to reduced variance within a subscale. In terms of relationships across measures of different constructs, labeling and pagination serve to separate measures of constructs (or dimensions of a construct); therefore, they may reduce observed correlations when compared to Condition 1.

In Condition 4, referred to as "interspersed," items from measures of different constructs were completely interspersed. Podsakoff et al. (2003) discusses the possibility of interspersion, referred to it as intermixing of items, which increases inter-construct correlation and reduces intra-construct correlation. Kline, Sulsky, and Rever-Moriyama (2000) include interspersion as a possible solution for reducing common method variance. Interspersion may detract from the halo effect within a measure of a construct when responses to later items are based on a general impression created by earlier items, likely reducing consistency across items within a measure representing a construct or a dimension. Interspersion can potentially cut both ways: it can serve to separate items within a construct, but can also create confusion and have the opposite effect of labeling or pagination. In this regard, the within-measure-of-a-construct halo effect serves to increase consistency of responses to items within a measure of a construct. Subsequent items in a construct are interpreted in light of earlier items. Researchers have shown increased reliability for later items in a construct (Knowles, 1988). However, interspersion also has the potential benefit of minimizing blurring across items from measures (subscales) representing different dimensions of a construct when compared to the contiguous condition, suggesting higher fit for multidimensional models with confirmatory factor analysis (CFA). Interspersion may also decrease observed relationships between measures of different constructs. This decrease when compared to Condition 1 may be greater than the decrease due to labeling or pagination.

In Condition 5, items were presented contiguously within measures (subscales) representing dimensions of the same construct as in Condition 1, but the sequence of measures/dimensions across the questionnaire was different. For example, two related measures of different constructs may be contiguous here but non-contiguous in Condition 1. Moreover, Condition 5 was sequenced so that no two measures (subscales) of dimensions of the same construct were contiguously placed, thereby testing the degree to which correlations between measures (subscales) representing dimensions of the same construct are influenced by contiguous placement by comparing to other conditions. Condition 6 was similar to Condition 5 with the addition of labeling which could lead to stronger relationships across related constructs when compared to Condition 5.

In Condition 7, referred to as "resequenced contiguous," items within subscales/measures of dimensions or constructs were sequenced differently when compared to the sequencing during their validation, but items of measures/subscales representing each construct or dimension were

still presented contiguously. The aim here was to assess the extent to which validated measures should be used with item sequencing identical to those at validation. The sequencing used at validation capitalizes on within-measure correlational systematic error due to sources such as a halo effect in responses. A different, unvalidated sequencing of items may detract from this effect. Condition 8, referred to as "resequenced, contiguous, and labeled," was similar to Condition 7 with the addition of labeling of measures (subscales) representing dimensions. This was designed to assess whether labeling would help overcome any detrimental effect due to resequencing of items.

Assessing the Impact of Questionnaire Format

A natural way to test whether various formatting conditions impact the psychometric properties of measures is multiple group confirmatory factor analysis (MGCFA; Steenkamp & Baumgartner, 1998). This procedure allows us to test whether measures are invariant across the conditions previously identified. Consider a vector of *k* items for the cth condition. The data model, x_c, is given by the following formula:

$$x_c = \tau_c + \Lambda_c \xi_c + \delta_c \quad (1)$$

where Λ_c is the matrix of factor loadings relating the vector of latent variables ξ_c to the observed items, τ_c is a vector of intercepts, and δ_c is a vector of measurement errors. Given a fixed condition, the covariance structure of the data Σ_c is computed as follows:

$$\Sigma_c = \Lambda_c \Phi_c \Lambda_c^T + \Theta_c \quad (2)$$

where Φ_c is the covariance matrix of the latent variables, Λ_c^T is the transpose of the matrix Λ_c and Θ_c is the covariance matrix of the error terms. Testing for measurement invariance essentially involves constraining specific parameters in equations (1) and (2) to be the same across conditions. Adopting the process suggested by Steenkamp and Baumgartner (1998), we impose increasingly more strict parameter constraints, in order to assess whether the psychometric properties of measures are invariant to the format used for collecting the data (Appendix B). Byrne, Shavelson, and Muthén (1989) suggest that there are two primary approaches to assessing invariance — measurement invariance and structural invariance, both of which encompass different components. The level of invariance required for a given model depends on the aims of the research. Measurement invariance encompasses configural, metric, scalar, and error variance invariance. *Configural* invariance exists when the patterns of loadings in the factor loading matrices ($\Lambda_1, \Lambda_2, \ldots, \Lambda_c$) indicate that the items in each condition load on to the same factor representing an underlying dimension. At the most basic level, the equivalence of factor loading structures implies that a construct can be conceptualized in the same way in terms of underlying dimensions under different conditions. It is tested by constraining factor loading structure to be identical across conditions, and comparing this constrained model to a benchmark model with the factor loading structures not constrained to be the same. More strict tests are for *metric* invariance, which occurs when the factor loadings are the same across conditions (i.e., $\Lambda_1 = \Lambda_2 = \ldots = \Lambda_c$), and for *scalar* invariance, which occurs when the intercepts are the same across conditions (i.e., $\tau_1 = \tau_2 = \ldots = \tau_c$). Once this level of invariance has been established, mean scores can be compared across conditions. *Error variance invariance* occurs when there is an approximately equal amount of measurement error across conditions (i.e., $\Theta_1 = \Theta_2 = \ldots = \Theta_c$), and along with metric invariance, suggests that a measure is equally reliable across conditions.

Structural invariance consists of factor variance and factor covariance invariance, and implies that the covariation or correlation between constructs is equivalent across conditions. *Factor variance invariance* is when the variance of the latent constructs is equal across conditions (i.e., $\phi_{i1} = \phi_{i2} = \ldots = \phi_{ic}$), whereas *factor covariance invariance* is when the latent constructs share the same covariance structure (i.e., $\phi_{ij1} = \phi_{ij2} = \ldots = \phi_{ijc}$ $i \neq j$).

The strength of MGCFA for our study is that it allows us to formally test whether questionnaire formatting impacts measures. MGCFA also allows us to formulate tests of format equivalence that are specifically appropriate for the different objectives of the research (Steenkamp & Baumgartner, 1998). When research aims to explore the structure of a construct and how various items relate to that construct, then lower levels of invariance such as configural and metric invariance may be sufficient. When the aim of research is to examine differences in mean scores between groups of individuals, then it may also be necessary to show scalar invariance. If, on the other hand, the aim of research is to test the relationships between constructs, as is often the case, higher levels of invariance are required. Researchers may wish to uncover the true scores of respondents without systematic bias. If measures depend on the format of the questionnaire rather than the underlying phenomena, different conclusions would be drawn, depending on the format used for collecting the data. The formal testing procedure is also particularly suited to our study since there is little theory to guide an understanding of the impact of different formats. The procedure used to test for invariance begins by allowing each condition to have a different model, and progressively applying each of the six constraints. In other words, in each of the eight conditions, six different invariance models were applied to the data and compared to establish which best explained the data. This allows us to identify where differences between conditions arise.

Following recommendations from the literature (e.g., Cheung & Rensvold, 2002; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000), we used standard statistics of goodness-of-fit for testing the six different invariance models. Whereas a standard chi-squared test of difference can be used to compare different models, such a test is extremely sensitive as the sample size increases, and is criticized as an impractical test of model fit (Cheung & Rensvold, 2002). Thus, we use root mean squared error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), and Bayesian information criterion (BIC) to determine whether a scale is invariant, with a far greater reliance on BIC as the most appropriate way to assess relative model fit.

Empirical Studies

We conducted five different studies to examine the impact of questionnaire formats on the psychometric properties of measures. Study 1 comprehensively tested all eight conditions across thirteen measures, using student samples. In Study 2, we examined whether an increase in cognitive demands on the respondent changes the pattern of results. In Study 3, we used stimulus-centered measures instead of respondent-centered measures that were used in Studies 1 and 2. In Study 4, we used stimulus-centered measures, but the correlation matrix for analysis was computed across stimuli, rather than across respondents (as in Study 3). Finally, in Study 5, we examined whether a non-student sample changes the pattern of results. Together, these five studies, eight conditions, and multitude of measures represent a large number of ways in which we sought to examine the implicit assumptions made by researchers about questionnaire formats. The five studies are linked

in testing different conditions in Study 1 with a student sample for respondent-centered measures, and extending to conditions of higher cognitive demand (Study 2), stimulus-centered measures (Studies 3 and 4), and a non-student sample (Study 5). Thus, we test our predictions for different study conditions, different types of measures, and different samples.

Study 1

Method

We tested questionnaire formats corresponding to the eight conditions (Appendix A), described previously, using a number of previously validated scales. The scales presented in Table 1 include coupon proneness, value consciousness, sale proneness, and price consciousness (Lichtenstein, Ridgway, & Netemeyer, 1993), consumer independent judgment-making and consumer novelty (Manning, Bearden, & Madden, 1995), consumer susceptibility to interpersonal influence (Bearden, Netemeyer, & Teel, 1989), material values — defining success, acquisition centrality, and pursuit of happiness (Richins & Dawson, 1992), consumer ethnocentrism (Shimp & Sharma, 1987), and need for cognition (Cacioppo, Petty, & Kao, 1984). We selected scales for which the response formats were similar to facilitate the presentation and analysis of various conditions described below. We used a test-retest approach with a 6-week interval. Data were collected from undergraduate students enrolled in introductory business classes at a large university with sample sizes in conditions ranging from 160-180. Students were given extra credit for participation. Each session took about 10-15 minutes to complete. Because of the sample sizes involved and the test-retest element, typically, data collection on one condition was completed during a semester.

TABLE 1
Scales and examples of items in studies

	Number of items	Sample items	Comment
		Studies 1 and 2	
Consumer independent judgment making (Manning et al., 1995)			
Consumer independent judgment making	6	Prior to purchasing a new brand, I prefer to consult a friend that has experience with the new brand	
Consumer novelty	8	I often seek out information about new products and brands	
Material values (Richins & Dawson, 1992)			
Defining success	6	I admire people who own expensive homes, cars, and clothes	
Acquisition centrality	7	I like a lot of luxury in my life	
Pursuit of happiness	5	My life would be better if I owned certain things I don't have	

(Table 1 continues)

Table 1 (continued)

	Number of items	Sample items	Comment
Value consciousness (Lichtenstein et al., 1993)			Also in Study 5
Value consciousness	7	I am very concerned about low prices, but I am equally concerned about product quality	
Price consciousness	5	I will grocery shop at more than one store to take advantage of low prices	
Coupon proneness	8	Redeeming coupons makes me feel good	
Sale proneness	5	If a product is on sale, that can be a reason for me to buy it	
Consumer ethnocentrism (Shimp & Sharma, 1987)	17	American people should always buy American-made products instead of imports	
Need for cognition (Cacioppo et al., 1984)	18	I find satisfaction in deliberating hard and for long hours	
Consumer susceptibility to interpersonal influence (Bearden et al., 1989)	12	I often consult other people to help choose the best alternative available from a product class	
		Study 3	
Service quality (Parasuraman, Zeithaml, & Berry, 1988)			
Tangibility	4	McDonald's has up-to-date equipment	
Reliability	5	What McDonald's promises to do something by a certain time, it does so	Also in Study 4
Responsiveness	4	You do not receive prompt service from McDonald's employees (R)	Also in Study 4
Assurance	4	You can trust the employees of McDonald's	
Empathy	5	McDonald's does not give you individual attention (R)	
Affective response to advertising (Holbrook & Batra, 1987)			
Pleasure	9	I felt grateful	
Arousal	9	I felt excited	
Domination	9	I felt afraid	
Endorser evaluation (Ohanian, 1990)		Please rate Michael Jordan as a celebrity endorser for Wheaties Cereal on the scales below	
Attractiveness	5	unattractive --- attractive	
Trustworthiness	5	undependable --- dependable	
Expertise	5	not an expert --- expert	

(Table 1 continues)

Table 1 (continued)

	Number of items	Sample items	Comment
Retail service quality (Dabholkar, Thorpe, & Rentz, 1996)			
Physical aspect	6	This store has modern-looking equipment	
Reliability	5	When this store promises to do something by a certain time, it will do so	
Personal interaction	9	Employees in this store give prompt service to customers	
Problem solving	3	This store willingly handles returns and exchanges	
Policy	5	This store offers high quality merchandise	
Perceived value			Also in Study 5
Perceived quality (Grewal, Monroe, & Krishnan, 1998)	3	The laptop appears to be of good quality	
Perceived transaction value (Grewal et al., 1998)	3	I would get a lot of pleasure knowing that I would save money at this reduced sale price	
Perceived acquisition value (Grewal et al., 1998)	9	This laptop would be a worthwhile acquisition because it would help me use it at a reasonable price	
Perceived sacrifice (Teas & Agarwal, 2000)	2	If I purchased the laptop for the indicated price, I would not be able to purchase some other products I would like to purchase now	
Perceived price (Zeithaml, 1988)	2	The price of this laptop is high	
Involvement (McQuarrie & Munson, 1986)			
Importance	5	The product is… important --- unimportant	
Interest	5	The product is… unexciting --- exciting	

Note. R = reverse coded.

Analysis and Results

A number of different types of analyses were conducted on the data to examine means and variances for subscales representing dimensions for each measure, item-to-total correlations and coefficient alphas for subscales representing dimensions for each measure, item level and overall test-retest correlations, exploratory and confirmatory factor analyses, and correlations across sub-scales representing dimensions and across measures of different constructs.

Means and Reliability

Means at test and at retest, as well as differences between means at test versus retest, did not suggest any consistent pattern across conditions. An examination of test-retest correlations across the conditions for each scale suggests no striking pattern across conditions that held for all measures (means of test-retest correlations across scales for each condition ranged from .68 to .75). Similarly, an examination of item level test-retest correlations across conditions suggests no striking pattern at this broad level (means of item level test-retest correlations across scales for each condition ranged from .49 to .57).

Similar results were found for item-to-total correlations and coefficient alpha. Means of item-to-total correlations across scales for each condition ranged from .54 to .76, with six conditions in the narrow band of .71-.76 and the continuous paginated condition and the measures resequenced labeled conditions in the 0.5 range. Means of coefficient alpha across scales for each condition ranged from .78 to .88 (see Table 2). Even interspersed versus labeling conditions did not lead to consistent differences across conditions (.84 versus .86, respectively), contrary to the assumption that the interspersed condition would lead to low internal consistency, at least for this particular sample and duration of administration. Lower internal consistency was found for some measures of scales in some conditions, such as material values — acquisition centrality scale in the paginated condition (coefficient alpha of .61 at test, for example). Similarly, item-to-total correlations were lower for some items of measures in some conditions. However, there was no dominant effect that held across conditions.

In order to establish whether these differences are statistically significant, we sought to conduct statistical tests to compare the variety of estimates of reliability across conditions. We note, however, that many of the estimates do not have well-defined distributions. Therefore, we use bootstrapping to construct 95% confidence intervals for all estimated statistics, described in Appendix C. The 95% confidence intervals were overlapping across all conditions and scales.

Factor Analyses

The next step in the analysis was to conduct factor analyses to examine the factor structure of various measures under different conditions. It is quite possible that differences across format conditions do not emerge in terms of means, internal consistency reliability, and stability reliability; yet, emerge in terms of blurring of items across subscales representing dimensions within and across measures. Indeed, internal consistency reliability assumes unidimensionality, which can be tested explicitly by factor analysis.

We began with exploratory factor analysis (EFA) and followed with confirmatory factor analysis (CFA), using appropriate multidimensional versus unidimensional models for measures of multidimensional and unidimensional constructs, respectively. Using CFAs, such an approach led to unsatisfactory levels of fit in all conditions. In an effort to boost to satisfactory levels at least in some conditions, we employed an approach involving parceling of items. We followed the rationale and recommendations for parceling suggested by Bagozzi and Edwards (1998, pp. 79-82) and Little, Cunningham, Golan, and Widaman (2002). We first used the results of EFA to select scales for further analyses where item loadings were high on appropriate subscales representing

TPM Vol. 24, No. 4, December 2017
465-498

TABLE 2
Study 1 – Test-retest correlations and coefficient alphas

Test-retest correlation across scales	Contiguous	Contiguous and labeled	Contiguous and paginated	Interspersed items	Resequenced measures	Resequenced and labeled measures	Resequenced items, contiguous measures	Resequenced items, contiguous and labeled measures
Consumer independent judgment making	.58	.50	.64	.54	.46	.44	.59	.60
Consumer novelty	.63	.66	.70	.70	.71	.57	.66	.68
Material values – Defining success	.79	.83	.75	.65	.82	.72	.83	.72
Material values – Acquisition centrality	.77	.73	.77	.77	.84	.70	.80	.73
Material values – Pursuit of happiness	.63	.74	.57	.76	.83	.72	.81	.78
Value consciousness	.69	.82	.59	.80	.77	.69	.78	.66
Price consciousness	.67	.72	.64	.61	.70	.69	.69	.68
Coupon proneness	.81	.73	.76	.80	.53	.67	.80	.62
Sale proneness	.70	.70	.68	.54	.68	.65	.76	.69
Consumer ethnocentrism	.78	.78	.78	.78	.81	.71	.78	.79
Need for cognition	.86	.85	.86	.79	.84	.78	.81	.82
Consumer susceptibility to interpersonal influence	.71	.75	.69	.73	.73	.75	.75	.70
Mean test-retest correlation	.72	.73	.70	.70	.73	.68	.75	.71

Coefficient alpha	test	retest	test	retest	test	retest	test	retest	Test	retest	test	retest	test	retest	test	retest
Consumer independent judgment making	.85	.91	.86	.89	.78	.80	.81	.86	.90	.89	.79	.63	.85	.89	.86	.89
Consumer novelty	.89	.91	.89	.91	.83	.85	.88	.90	.92	.91	.86	.69	.88	.92	.89	.88
Material values – Defining success	.85	.89	.84	.84	.73	.77	.80	.54	.84	.88	.75	.76	.86	.87	.83	.84
Material values – Acquisition centrality	.81	.82	.77	.76	.61	.65	.75	.82	.81	.81	.64	.67	.76	.82	.74	.73
Material values – Pursuit of happiness	.76	.75	.75	.84	.65	.43	.83	.83	.83	.84	.71	.64	.77	.84	.83	.84
Value consciousness	.84	.89	.87	.87	.84	.86	.82	.85	.88	.85	.88	.88	.85	.86	.85	.87
Price consciousness	.78	.85	.80	.80	.71	.66	.76	.83	.82	.83	.99	.71	.82	.83	.80	.84

(Table 2 continues)

Table 2 (continued)

Coefficient alpha	test	retest	test	retest	test	retest	test	retest	test	retest	test	retest	test	retest	test	retest
Coupon proneness	.90	.93	.91	.91	.88	.89	.89	.90	.84	.56	.84	.84	.90	.90	.86	.87
Sale proneness	.79	.82	.84	.78	.84	.83	.81	.83	.85	.84	.82	.82	.83	.86	.83	.83
Consumer ethnocentrism	.96	.97	.96	.97	.95	.95	.96	.96	.93	.96	.97	.95	.97	.97	.97	.97
Need for cognition	.90	.92	.90	.92	.84	.81	.89	.92	.88	.93	.79	.84	.90	.84	.89	.92
Consumer susceptibility to interpersonal influence	.78	.91	.87	.90	.89	.89	.86	.88	.89	.91	.90	.91	.87	.89	.89	.90
Mean coefficient alpha	.84	.88	.86	.87	.79	.78	.84	.84	.87	.85	.81	.78	.86	.87	.86	.87

dimensions (≥ .40) and not high on inappropriate subscales representing dimensions (< .25). By this criterion, we excluded the need for cognition scale and the susceptibility to interpersonal influence scale from further analysis, which did not yield interpretable factor structures according to specifications for the original scales. We then used a parceling approach that was not based on idiosyncratic item content and that did not vary across scales, essentially using generic rules for combining contiguous items into different parcels. For subscales representing dimensions with relatively few items, two parcels for each subscale representing a dimension were used comprised of odd versus even numbered items, respectively. For the longer ethnocentrism scale, we used four parcels, each consisting of every fourth item with a different starting point (e.g., Item 1, 5, 9 … in the first parcel; 2, 6 … in parcel 2, etc.).

Having established measurement models, we conducted MGCFA to test for invariance. For each scale, we estimated three sets of MGCFAs — a model combining test and retest, a model for test alone, and a model for retest alone. Results are presented in Table 3. We discuss the first case for the "test" set of the Materialism scale. We began by testing for configural invariance, which is when the factor loading structure is the same across conditions. This is considered appropriate since the model has acceptable fit (RMSEA = .071; CFI = .991; TLI = .970). Testing for metric invariance, the results support the conclusion that the factor loadings are invariant across conditions (RMSEA = .051; CFI = .993; TLI = .985); and we find similar support for scalar invariance (RMSEA = .074; CFI = .979; TLI = .968). This finding implies that if the aim of a study was to compare means, it would not matter which of the questionnaire formats were implemented. We find support for factor variance invariance (RMSEA = .073; CFI = .963; TLI = .968) and for factor covariance (RMSEA = .078; CFI = .967; TLI = .964). These results suggest that choice of questionnaire format will not lead to different conclusions when examining relationships between constructs. We also find support for error variance invariance (RMSEA = .074; CFI = .951; TLI = .968). Along with metric invariance, this result suggests that the eight conditions are similarly reliable, reinforcing the previous findings. Further support for measurement invariance is provided by the BIC, which is minimized at the final level of invariance (35247.34).

We obtained similar results for the remainder of the measures. 12 out of 12 MGCFAs (three sets for each of the four scales), achieved all six levels of invariance using goodness-of-fit criteria. Furthermore, in 11 of the cases the BIC was minimized in the most restrictive invariance model (all six levels; the one exception was the retest set of the consumer independent judgment and decision making scale, which did achieve the first five levels of invariance), suggesting consistence in the overall pattern of results.

Cross-Dimensional Relationships

To further examine the impact of questionnaire format on the relationships between dimensions within a construct, we computed the correlations among the set of dimensions or constructs within each of three multidimensional construct measures or subscales (Table 4A). Again, no discernible differences were found across conditions, with bootstrapping yielding an overwhelming pattern of overlapping confidence intervals. Our earlier discussion of the motivations for different formats argued for the possibility of labeling separating each distinct measure (subscale) of individual dimensions in respondents' minds, thus lowering the observed relationship but

TABLE 3
Study 1 – Multiple group confirmatory factor analysis

	Invariance	Test				Retest				Test-Retest			
Measure	Type	RMSEA	CFI	TLI	BIC	RMSEA	CFI	TLI	BIC	RMSEA	CFI	TLI	BIC
Material values	Configural	.071	.991	.970	35764.12	.066	.993	.978	35584.09	.128	.941	.879	68652.81
	Metric	.051	.993	.985	35673.71	.078	.985	.969	35518.58	.121	.939	.892	68496.23
	Scalar	.074	.979	.968	35617.79	.093	.972	.956	35467.68	.121	.930	.892	68391.41
	Factor covariance	.078	.967	.964	35514.32	.094	.959	.955	35371.82	.105	.924	.918	67795.99
	Factor variance	.073	.963	.968	35393.30	.094	.947	.955	35272.83	.104	.916	.920	67601.92
	Error variance	.074	.951	.968	35247.34	.092	.934	.957	35133.07	.101	.909	.925	67308.49
Consumer ethnocentrism	Configural	.233	.978	.935	25865.55	.222	.981	.942	25880.06	.174	.949	.925	49958.80
	Metric	.163	.975	.968	25755.90	.155	.978	.971	25766.78	.156	.948	.940	49720.51
	Scalar	.162	.962	.968	25715.01	.133	.975	.979	25658.69	.151	.941	.944	49567.87
	Factor covariance	.162	.962	.968	25715.01	.133	.975	.979	25658.69	.149	.940	.945	49528.65
	Factor variance	.157	.960	.970	25685.28	.127	.974	.981	25620.03	.148	.938	.946	49470.30
	Error variance	.156	.943	.971	25622.15	.117	.968	.984	25485.14	.144	.929	.949	49267.45
Consumer independent judgment making	Configural	.101	.995	.968	26989.69	.196	.982	.890	27011.37	.096	.975	.949	52517.70
	Metric	.167	.960	.913	26991.33	.216	.939	.867	27043.86	.086	.974	.959	52346.44
	Scalar	.057	.992	.990	26822.22	.174	.935	.914	26967.61	.090	.967	.956	52224.91
	Factor covariance	.057	.991	.990	26782.22	.161	.934	.926	26928.33	.082	.965	.963	51974.67
	Factor variance	.070	.982	.985	26718.63	.081	.978	.981	26718.04	.088	.955	.958	51871.22
	Error variance	.104	.940	.966	26652.03	.140	.902	.944	26758.98	.101	.926	.944	51715.86
Value consciousness	Configural	.068	.986	.967	49136.09	.067	.988	.971	48046.10	.111	.930	.875	94741.07
	Metric	.073	.979	.962	49032.59	.064	.985	.974	47925.79	.108	.928	.883	94503.70
	Scalar	.108	.943	.916	49045.04	.094	.962	.944	47900.31	.116	.908	.864	94483.88
	Factor covariance	.094	.942	.937	48789.34	.086	.956	.953	47664.39	.102	.901	.895	93351.34
	Factor variance	.092	.933	.939	48649.97	.092	.942	.947	47555.58	.102	.894	.896	93094.90
	Error variance	.092	.918	.939	48449.59	.089	.932	.949	47338.54	.101	.884	.898	92693.79

Note. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; BIC = Bayesian information criterion.

TABLE 4A
Average cross-dimensional correlations

		Contiguous	Contiguous and labeled	Contiguous and paginated	Interspersed items	Resequenced measures	Resequenced and labeled measures	Resequenced items, contiguous measures	Resequenced items, contiguous and labeled measures
		N = 183	*N* = 165	*N* = 160	*N* = 180	*N* = 161	*N* = 170	*N* = 170	*N* = 159
Study 1 – test	Consumer novelty	–.36	–.13	–.35	–.15	–.25	–.33	–.20	–.27
	Materialism	.46	.38	.39	.44	.58	.37	.55	.47
	Value consciousness	.34	.31	.31	.43	.48	.39	.45	.43
Study 1 – retest	Consumer novelty	–.27	–.32	–.48	–.30	–.33	.00	–.29	–.25
	Materialism	.51	.42	.38	.40	.60	.37	.57	.56
	Value consciousness	.39	.40	.39	.44	.45	.39	.48	.51
Study 2	Consumer novelty	–.10	.00		–.29				
	Materialism	.46	.43		.56				
	Value consciousness	.44	.33		.50				
Study 3	Service quality	.48	.38		.52				
	Affect	.31	.27		.22				
	Endorsement	.48	.56		.74				
	Retail service quality	.46	.49		.46				
	Computer evaluation	.25	.31		.35				
	Involvement	.62	.48		.65				
Study 5	Value consciousness		–.02		.10				
	Computer evaluation		.32		.29				

not with contiguous or interspersed conditions. Whereas interspersion serves to separate items, labeling may serve to group items within a subscale representing a dimension and distinguish them from items from other subscales representing other dimensions, and attenuate relationships. However, the results did not support these assumptions, with no consistent pattern indicating that a particular format impacts the strength or direction of relationships between dimensions. The measures resequenced conditions where subscales representing dimensions were not presented contiguously did not differ from the other conditions.

Cross-Construct Relationships

Next, correlations across constructs were examined. In particular, the relationship between consumer independent judgment making and susceptibility to interpersonal influence was examined. We expected a negative relationship between these two constructs at a conceptual level, independence being likely to be negatively related to susceptibility to influence. All the items in the former measure related to seeking advice or consulting friends before purchase. Several items in the susceptibility to interpersonal influence relate to seeking advice from friends. In the measures resequenced conditions, consumer independent judgment making and susceptibility to interpersonal influence were presented contiguously. Whereas correlations were statistically significant across most conditions, there were no consistent differences across conditions (Table 4B), as assessed with a *z*-test. For example, the difference between correlations in the contiguous and contiguous labeled for the "test" set is significant (–.37 vs. –.03, $p < 0.01$), whereas the same difference becomes insignificant for the "retest" set (–.36 vs. –.37). A similar inconsistent pattern was found with correlations of susceptibility to interpersonal influence and consumer novelty.

Discussion

The overwhelmingly striking and perhaps surprising pattern from Study 1 was the invariance of measures across conditions for a student sample. This implies that one could conduct theory testing by comparing subsets of individuals, or examining relationships between variables, and regardless of the format of the questionnaire, the substantive conclusions would be similar. Additionally, it appears that, irrespective of condition, measures performed comparably across conditions in terms of test-retest reliability and internal consistency reliability. These results indicate that our study does not provide evidence of the impact of questionnaire format on the psychometric properties of the scale. Therefore, the rationale leading to predictions of the impact of format differences did not find support in our data.

In interpreting these findings, even the interspersed condition was not different from the labeled condition. In other words, in one condition, items are disguised by interspersing with other items, and in the other, items are labeled under a subscale representing a dimension or a scale representing a unidimensional construct and presented together. On the one hand, this pattern points to the quality of the measures used. However, this pattern may also be a consequence of the administration in Study 1, which involved completion of these measures by students for a length

TABLE 4B
Average cross-construct correlations

		Contiguous	Contiguous and labeled	Contiguous and paginated	Interspersed items	Resequenced measures	Resequenced and labeled measures	Resequenced items, contiguous measures	Resequenced items, contiguous and labeled measures
		N = 183	*N* = 165	*N* = 160	*N* = 180	*N* = 161	*N* = 170	*N* = 170	*N* = 159
Study 1 – test	SUS-CON	–.38	–.03	–.44	–.37	–.20	–.20	–.15	–.27
	SUS-NOV	.28	.35	.28	.30	.29	.31	.14	.09
Study 1 – retest	SUS-CON	–.36	–.37	–.37	–.41	–.20	–.08	–.32	–.33
	SUS-NOV	.38	.38	.25	.26	.42	.19	.27	.21
Study 2	SUS-CON	–.50	–.20		–.36				
	SUS-NOV	.38	.12		.22				
Study 3	PTRAN-PPRI	.27	.32		.53				
	PACV-PPRI	.35	.54		.68				
	PTRAN-PQUAL	.54	.39		.53				
	PACV-PQUAL	.66	.52		.73				
	PTRAN-PSAC	–.17	.06		–.11				
	PACV-PSAC	–.16	.09		.01				
	PPRI-PQUAL	.16	.19		.35				
	PPRI-PSAC	.26	.27		.16				
Study 5	PTRAN-PPRI		.09		.32				
	PACV-PPRI		–.06		.07				
	PTRAN-PQUAL		.62		.55				
	PACV-PQUAL		.67		.72				
	PTRAN-PSAC		.40		.20				
	PACV-PSAC		.25		–.05				
	PPRI-PQUAL		.18		.15				
	PPRI-PSAC		.11		.39				

Note. SUS-CON = Consumer susceptibility to interpersonal influence-Consumer independent judgment making; SUS-NOV = Consumer susceptibility to interpersonal influence-Consumer novelty; PTRAN-PPRI = Perceived transaction value-Perceived price; PACV-PPRI = Perceived acquisition value-Perceived price; PTRAN-PQUAL = Perceived transaction value-Perceived quality; PACV-PQUAL = Perceived acquisition value-Perceived quality; PTRAN-PSAC = Perceived transaction value-Perceived sacrifice; PACV-PSAC = Perceived acquisition value-Perceived sacrifice; PPRI-PQUAL = Perceived price-Perceived quality; PPRI-PSAC = Perceived price-Perceived sacrifice.

of time that was typically between 10 and 15 minutes. In other words, the demands on the respondents were not burdensome in terms of administration procedures. We use Study 2 to examine whether a greater burden on respondents might generate a different pattern of results.

Study 2

Method

To disentangle the effect of the overall cognitive demands of the administration and examine its role as a potential moderating factor, a second study was conducted. A student sample completed the same questionnaire used in Study 1 as a part of data collection of approximately three times the duration (about 45 minutes). The relationship between questionnaire length and response quality has been investigated in the methodological literature (Burchell & Marsh, 1992; Herzog & Bachman, 1981; Krosnick 1999), and recommendations on this issue are common in survey research. Students were assigned to one of three conditions considered the most distinctly different: the contiguous, the labeled, and the interspersed conditions (Appendix A). Whereas the labeled provides a sharp logical division, the interspersed is diametrically opposite; together these two conditions represent the most distinctly different conditions. The contiguous condition provides a moderate baseline for comparison. There was a test phase but no retest phase in this study. Students completed the same measures as in Study 1, but as part of a larger study that typically took about 45 minutes to complete. The questionnaire of interest for this study took about 15 minutes, as in Study 1, and was typically administered at the middle or the end of the administration of the larger study.[1] The sample size ranged from 129 to 131 for each condition.

Analysis and Results

There were no sizable differences in means or standard deviations between the three conditions. In terms of internal consistency, coefficient alphas were similar in magnitude across conditions (.81 to .84). In some instances, the interspersed condition had lower item-to-total correlations (e.g., value consciousness: .80 for interspersed vs. .86 for labeled and .89 for contiguous). However, this was an exception rather than evidence for an overwhelming or a consistent pattern. When compared to Study 1, where the cognitive demands were lower, lower coefficient alphas were found only for the need for cognition scale. Once again, most bootstrapped 95% confidence intervals were overlapping, and where there was no overlap, there was no pattern.

Using procedures identical to those in Study 1, items were parceled for each of the subscales representing dimensions. There were also no consistent differences across the conditions in terms of MGCF (Table 5). As in Study 1, the CFI and the TLI supported invariance across conditions, while the BIC preferred the model with all six levels of invariance imposed. The only exception was for the consumer judgment scale, which had poor fit across all the conditions.

Another set of findings related to relationships across dimensions within a construct. Correlations were computed among the set of dimensions or constructs within each of three multidimensional measures. No consistent difference emerged across conditions (Table 4A). Correlations across

constructs were also examined and did not lead to any striking differences across conditions (Table 4B). These results indicate that our data do not provide evidence of differences across conditions.

TABLE 5
Study 2 – Multiple group confirmatory factor analysis

Measure	Invariance type	RMSEA	CFI	TLI	BIC
Material values	Configural	.000	1.000	1.006	10057.99
	Metric	.011	1.000	.999	10039.58
	Scalar	.040	.995	.991	10023.06
	Factor covariance	.018	.999	.998	9990.703
	Factor variance	.020	.998	.998	9961.434
	Error variance	.000	1.000	1.001	9919.028
Consumer ethnocentrism	Configural	.137	.992	.975	7884.139
	Metric	.099	.991	.987	7855.016
	Scalar	.112	.984	.984	7839.252
	Factor covariance	.112	.984	.984	7839.252
	Factor variance	.106	.984	.985	7829.319
	Error variance	.084	.986	.991	7785.621
Consumer independent judgment making	Configural	.344	.945	.668	8894.712
	Metric	.237	.939	.842	8879.913
	Scalar	.187	.940	.902	8858.877
	Factor covariance	.180	.934	.909	8853.658
	Factor variance	.171	.922	.918	8843.898
	Error variance	.150	.912	.936	8812.934
Value consciousness	Configural	.058	.991	.979	15593.56
	Metric	.069	.985	.971	15572.28
	Scalar	.084	.974	.958	15557.78
	Factor covariance	.067	.978	.973	15491.76
	Factor variance	.085	.960	.957	15477.55
	Error variance	.094	.940	.947	15444.44

Note. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; BIC = Bayesian information criterion.

Discussion

The results of Study 2 suggest that cognitive load arising from the duration of data collection procedures did not contribute to the lack of differences observed in Study 1. The same pattern of results in terms of measurement invariance, cross-dimensional relationships, and cross-construct relationships were observed as in Study 1. It appears that, irrespective of a diverse set of conditions, student respondents are able to respond appropriately based on the content of items, even when the duration of data collection is extended. Compared to others, students may be highly motivated to fill-out questionnaires, more familiar with, and practiced in filling out questionnaires, and more cognitively complex or skilled in processing abstract concepts often comprising questionnaires.

Study 3

Study 3 was conducted to examine an important factor that may moderate the results, the nature of the measures on which data are collected. The measures used in earlier studies are respondent-centered, relating to individual differences, rather than stimulus-centered or relating to characteristics of stimuli (Cox, 1980). The nature of the measures possibly overwhelms the effects of format factors. In particular, respondents may be more certain and knowledgeable about traits and characteristics pertaining to themselves. Literature from a number of areas of research including self-concepts, self-referencing, and autobiographical memory (e.g., Krishnamurthy & Sujan, 1999) supports this line of reasoning, arising from the highly-organized memory structure of the self (Greenwald & Banaji, 1989) and leading to advantages in elaboration of incoming information and memory. When using validated scales with items relating to the self, student respondents are perhaps able to complete questionnaires based on item content, minimizing the effect of format factors. Therefore, a study was designed using stimulus-centered scales under the three conditions employed in Study 2 (Appendix A). The rationale for choosing these three conditions was discussed earlier; that is, comparing the most distinctly different labeling and interspersed conditions along with the baseline contiguous condition. Because each set of measures relating to a construct pertained to a different stimulus, interspersion was carried out between items from subscales of different dimensions of a construct rather than across constructs.

Method

A range of stimulus-centered scales was used in the study including measures of multidimensional constructs: service quality (Parasuraman et al., 1988) for McDonalds fast food restaurants with five dimensions and associated subscales — tangibility (four items), reliability (five items), responsiveness (four items), assurance (four items), and empathy (five items); affective response to an ad (Holbrook & Batra, 1987) for a health club with three dimensions and associated subscales — pleasure (six items), arousal (six items), and dominance (six items); perceived expertise (five items), attractiveness (five items), and trustworthiness (five items) of a celebrity endorser (Michael Jordan on a Wheaties cereal box; Ohanian, 1990); retail service quality (Dabholkar et al., 1996) of Walmart stores with five dimensions and associated subscales — physical aspects (six items), reliability (five items), personal interaction (nine items), problem solving (three items), and policy (five items); evaluation of a computer based on a picture and description on perceived quality (three items; Grewal et al., 1998), perceived transaction value (three items; Grewal et al., 1998), perceived acquisition value (nine items; Grewal et al., 1998), perceived sacrifice (two items; Teas & Agarwal, 2000), and perceived price (two items adapted from multiple sources; construct discussed in Zeithaml, 1988, and other literature); involvement (McQuarrie & Munson, 1986) in a smart phone based on a picture and description with two dimensions and associated subscales — importance (five items) and interest (five items). Stimuli were generally selected to be moderately positive to allow for variation on the scales. The sample was again made up of students, with the sample size ranging from 127 to 131 for each condition.

Analysis and Results

Means and standard deviations were comparable across conditions. The interspersed condition had slightly lower mean coefficient alphas across scales (.75) when compared to the contiguous and contiguous labeled conditions (.82 and .82), with this pattern being accentuated for some dimensions (see Table 6; e.g., dominance dimension of affect: .70 for contiguous labeled vs. .48 for interspersed). For stimulus-centered scales, internal consistency may be somewhat lower for interspersed items, with interspersion leading to separation between individual items within a subscale capturing a dimension. On the other hand, this effect does not occur for contiguous and labeled conditions. The bootstrapped 95% confidence intervals were overlapping for the most part, with no consistent pattern.

TABLE 6
Study 3 – Alpha coefficients

		Contiguous	Contiguous and labeled	Interspersed items
Service quality (McDonald's)	Tangibility	.69	.68	.59
	Reliability	.82	.80	.74
	Responsiveness	.77	.68	.69
	Assurance	.75	.77	.70
	Empathy	.77	.71	.64
	Mean	.76	.73	.67
Affective response to ad (Print ad/health club)	Pleasure	.87	.89	.79
	Arousal	.75	.76	.63
	Dominance	.60	.70	.48
	Mean	.74	.78	.64
Endorser evaluation (Michael Jordan for Wheaties cereal)	Attractiveness	.83	.86	.83
	Trustworthiness	.87	.92	.87
	Expertise	.90	.92	.91
	Mean	.87	.90	.87
Retail service quality (Walmart)	Physical aspect	.78	.78	.73
	Reliability	.84	.84	.74
	Personal interaction	.89	.89	.84
	Problem solving	.77	.85	.74
	Policy	.53	.47	.21
	Mean	.76	.77	.65
Perceived value (Computer)	Perceived quality	.90	.91	.79
	Perceived transaction value	.87	.89	.92
	Perceived acquisition value	.97	.97	.96
	Perceived sacrifice	.86	.86	.83
	Perceived price	.86	.92	.78
	Mean	.89	.91	.86
Involvement (PalmOne)	Importance	.94	.93	.88
	Interest	.94	.91	.91
	Mean	.94	.92	.90
Mean		.82	.82	.75

In terms of MGCFA, odd-even parceling was used as in previous studies. Again, almost all of the measures met the criteria of invariance, with the CFI greater than .92 for all models, the TLI greater than .93 for all models, and the BIC minimized when all six forms of invariance imposed (Table 7). The RMSEA was acceptable for all models with the only exception being for the involvement construct, which had a RMSEA < .10 when testing for factor variance invariance. Despite this, there were again no overwhelming differences in patterns of results that held across conditions.

TABLE 7
Study 3 – Multiple group confirmatory factor analysis

Measure	Invariance type	RMSEA	CFI	TLI	BIC
Service quality (McDonald's)	Configural	.059	.989	.967	11820.58
	Metric	.048	.991	.978	11798.86
	Scalar	.065	.981	.960	11787.44
	Factor covariance	.057	.976	.969	11692.84
	Factor variance	.067	.961	.958	11656.52
	Error variance	.083	.931	.936	11642.01
Affective response to ad (Print ad/health club)	Configural	.095	.975	.938	11256.08
	Metric	.082	.975	.953	11226.48
	Scalar	.098	.956	.934	11212.91
	Factor covariance	.088	.957	.946	11182.23
	Factor variance	.092	.945	.941	11162.46
	Error variance	.094	.926	.939	11118.68
Endorser evaluation (Michael Jordan for Wheaties cereal)	Configural	.046	.997	.993	10426.44
	Metric	.042	.997	.994	10397.26
	Scalar	.053	.993	.990	10372.98
	Factor covariance	.057	.991	.989	10347.22
	Factor variance	.093	.972	.970	10349.47
	Error variance	.095	.963	.969	10305.73
Retail service quality (Walmart)	Configural	.065	.974	.959	15507.18
	Metric	.064	.972	.960	15452.82
	Scalar	.068	.964	.953	15409.58
	Factor covariance	.064	.964	.960	15310.54
	Factor variance	.073	.951	.948	15290.44
	Error variance	.074	.942	.946	15199.55
Perceived value (Computer)	Configural	.073	.975	.963	16135.9
	Metric	.070	.975	.966	16066.42
	Scalar	.069	.973	.966	16004.98
	Factor covariance	.067	.972	.969	15909.01
	Factor variance	.078	.960	.958	15905.09
	Error variance	.091	.938	.942	15866.13
Involvement (PalmOne)	Configural	.062	.999	.994	6840.614
	Metric	.052	.998	.996	6821.706
	Scalar	.068	.996	.993	6805.935
	Factor covariance	.080	.993	.990	6800.328
	Factor variance	.116	.980	.978	6799.402
	Error variance	.127	.964	.974	6781.989

Note. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; BIC = Bayesian information criterion.

Another set of findings related to relationships across subscales of dimensions within a construct and across measures of different constructs. Although not universally the case, there were instances of lower correlations for the labeled condition when compared to the contiguous or interspersed conditions (e.g., Tables 4A and 4B). This may be due to the logical division of a dimension or a construct that labeling may achieve, as discussed earlier.

Discussion

Overall, Study 3 suggests that the nature of scales (stimulus-centered vs. respondent-centered) is not a significant factor in moderating the effects of format factors on responses for a student sample. Therefore, the lack of differences across conditions is not due to the nature of items pertaining to the self.

Study 4

Method

To further examine stimulus-centered scales, Study 4 used a design where correlations were computed on ratings *across stimuli* rather than respondents, again comparing the labeled, interspersed, and contiguous conditions (Appendix A), the rationale for choosing these three conditions having been discussed earlier. For each condition, five versions were created. Each version involved approximately 20 respondents rating a total of 20 restaurants. Each restaurant was rated on two dimensions (each subscale capturing a dimension with three items): reliability and responsiveness. The aim here was to keep the length of the questionnaire comparable to the lengths in the first three studies, yet collect data on multiple dimensions of service quality. Three items from measures (subscales) of each dimension were used leading to a total of 120 items (six items each for 20 restaurants). Across five versions, the total number of restaurants rated was 100, which becomes the effective sample size. Means were computed for each restaurant on each item. Correlations were then computed across restaurants, with a sample size of 100 providing a basis for MGCFA.

Analysis and Results

The results did not suggest any striking differences across the three conditions (mean coefficient alphas across conditions ranged from .86 to .89; Table 8). The bootstrapped 95% confidence intervals do not display a consistent pattern, with most intervals overlapping.

As there were only three items for each subscale per dimension, parceling was not used in specifying the MGCFA models. Although the RMSEA was high in this study, the CFI, TLI, and BIC all provided evidence for invariance across formatting conditions (Table 9), suggesting that differences across conditions do not emerge for analyses based on correlations across stimuli rather than individuals. For student samples used in our studies, the nature of the analysis, that is, across stimuli versus individuals, does not lead to differences. The high RMSEA is to be expected with the sample size of 20 respondents (as RMSEA does not account for the multiple stimuli being

responded to by these 20 respondents). The BIC is more appropriate here, which provides clear evidence of invariance.

TABLE 8
Alpha coefficients for Study 4

Alpha coefficients	Contiguous and labeled	Contiguous	Interspersed items
Reliability	.91	.92	.91
Responsiveness	.87	.80	.80
Mean	.89	.86	.86

TABLE 9
Study 4 – Multiple group confirmatory factor analysis

Measure	Invariance type	RMSEA	CFI	TLI	BIC
Service quality	Configural	.209	.930	.869	17121.39
	Metric	.184	.929	.900	17086.47
	Scalar	.169	.924	.915	17055.47
	Factor covariance	.164	.925	.920	17044.90
	Factor variance	.160	.922	.924	17029.95
	Error variance	.145	.919	.937	16978.53

Note. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; BIC = Bayesian information criterion.

Discussion

Overall, Study 4 suggests that the nature of scales (stimulus-centered vs. respondent-centered) using stimuli as the unit of analysis is not a significant factor in moderating the effects of format factors on responses for a student sample.

Study 5

Method

Study 5 was conducted to examine the effect of another factor, the sample composition, by using a non-student adult sample. The first four studies employed student samples, producing few, if any, striking differences across conditions and thus demonstrating the robustness of published scales to variations in format. Stimulus- versus respondent-centered scales also did not lead to significant differences across conditions. The aim in this study was to examine whether the lack of

differences across conditions, particularly for the respondent-centered measures, is because of students' skills in taking questionnaires and frequent participation in studies, as well as their relatively high cognitive skills and tolerance for abstractions. In this regard, Churchill and Peter (1984) hypothesized but did not find results supportive of the notion that student samples lead to higher reliability. Peterson (2001) conducted a meta-analysis that identified differences in homogeneity and effect sizes between student and non-student samples. Research across several disciplines has examined this issue (e.g., Gordon, Slade, & Schmitt, 1986).

In Study 5, two conditions — labeled and interspersed, the two most distinctly different as discussed earlier — were compared (Appendix A), using a subset of respondent- and stimulus-centered scales (respondent-centered scales: coupon proneness, value consciousness, sale proneness, and price consciousness; stimulus-centered scales, identical to Study 3: evaluation of a computer based on a picture and description on perceived quality, perceived transaction value, perceived acquisition value, perceived sacrifice, and perceived price). Non-student adults were recruited in a university town through several means; by having volunteers at a local non-profit organization complete the questionnaire and by approaching employees of a large university. Questionnaires were distributed through supervisors who asked participants to complete them. One hundred and seventy-four individuals participated in this study, about equally distributed across the two conditions.

Analysis and Results

Analyses similar to previous studies were conducted. Whereas means and standard deviations were largely similar, differences in coefficient alpha emerged for some of the stimulus-centered scales, specifically, perceived transaction value and perceived price scales (Table 10).

TABLE 10
Alpha coefficients for Study 5

	Contiguous and labeled	Interspersed items
Value consciousness	.84	.78
Price consciousness	.56	.71
Coupon proneness	.90	.83
Sale proneness	.71	.70
Mean	.75	.76
Perceived quality	.89	.72
Perceived transaction value	.84	.58
Perceived acquisition value	.93	.86
Perceived sacrifice	.82	.79
Perceived price	.57	.18
Mean	.81	.63
Overall mean	.78	.68

Although the overall mean coefficient alphas (of .78 vs. .68) were somewhat different across conditions, these differences were driven by one construct (the perceived price scale) and should therefore be viewed with caution. As in all other studies, the bootstrapped 95% confidence intervals indicate no pattern, and are overlapping for most conditions.

In terms of MGCFA, the respondent-centered value consciousness scales failed to achieve reasonable fit even when testing for configural invariance again. This is largely due to the problematic measure of the perceived price dimension (Table 11). In particular, the interspersed condition for value consciousness led to an unsatisfactory level of fit in contrast to the contiguous labeled condition, the only divergence between conditions in terms of CFA results that we observed across our five studies. The stimulus-centered computer evaluation scales achieved factor covariance invariance, but not variance invariance or error variance invariance.

TABLE 11
Study 5 – Multiple group confirmatory factor analysis

Measure	Invariance type	RMSEA	CFI	TLI	BIC
Service quality (McDonald's)	Configural	.119	.943	.867	6549.633
	Metric	.153	.890	.780	6557.791
	Scalar	.170	.846	.730	6562.632
	Factor covariance	.151	.847	.786	6537.251
	Factor variance	.141	.849	.813	6519.832
	Error variance	.128	.854	.846	6493.026
Affective response to ad (Print ad/health club)	Configural	.088	.943	.914	6969.076
	Metric	.087	.940	.917	6942.218
	Scalar	.093	.927	.905	6927.056
	Factor covariance	.091	.922	.908	6890.448
	Factor variance	.103	.897	.884	6895.734
	Error variance	.117	.852	.849	6892.397

Note. RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; BIC = Bayesian information criterion.

Discussion

Overall, Study 5 suggests that, for the most part, the nature of the sample composition is not a significant factor in moderating the effects of format factors on responses. However, some differences emerged such as the finding that the interspersed condition for the value consciousness scale led to a poor fit relative to the labeled condition.

GENERAL DISCUSSION

We set out to investigate whether and how different format practices affect research outcomes, and, in particular, the psychometric properties of measures and the observed relationships between constructs. Researchers use a variety of formats, reflecting implicit beliefs of how each of these practices ameliorates a measurement-related problem. The surprising and consistent findings

from our research suggest that implicit theories about labeling, interspersion and the like may not necessarily hold for student samples. By and large, we do not find differential effects of such presentation issues on measure reliability and validity. In fact, magnitudes of means and standard deviations, internal consistency reliability, stability reliability, dimensionality, cross-dimensional relationships, and cross-construct relationships are, for the most part, unaffected by such variations in format. In particular, formal statistical tests for measurement invariance support the view that formatting for the most part does not impact the psychometric properties of measures for the samples we employed. Arguments such as a possible grouping effect created by labeling versus the opposite created by interspersion do not hold up under empirical testing. The 95% confidence intervals overlap for the most part, an overwhelming pattern further indicating that the differences are statistically insignificant.

Where labeling and interspersion may have an effect, though, is perhaps with non-student samples and respondent-centered scales. Whereas differences across conditions are minimal for student samples, differences emerge for non-student samples. Specifically, we find that the interspersed condition for the value consciousness scale led to a poor fit relative to the labeled condition when we used a non-student sample. However, this difference must be put in the context of the large number of ways in which we explored possible effects across a variety of conditions, scales, and samples. In Study 1, we used 12 scales, compared eight formats for those scales, and found no differences of consequence. In Study 2 (higher cognitive burden relative to Study 1), we used four scales, compared three formats, and found no differences. In Study 3 (stimulus-centered scales vs. respondent-centered scales), we used six scales, compared three formats, and found no differences. In Study 4 (stimulus as unit of analysis), we used one stimulus-centered scale, compared three conditions, and again found no differences. In Study 5 (non-student sample), we used a mix of nine stimulus- and respondent-centered scales, compared two formats, and found a difference in fit across conditions for only one scale. In summary, across a large number of tests, we found only one difference, although only comparing two conditions in this study. This difference though, emphasizes the need for further research on non-student samples across a variety of conditions. Futhermore, there is a need for further research to understand effects at the level of specific measures.

Our research questions the implicit and explicit rationales about presentation effects. By and large, we did not find empirical support for presumed differences between such presentation factors as labeling and interspersion. Thus, when such presentation factors are suspected a priori, either empirical support — often impractical in substantive studies as it would involve experimental manipulation of methodological factors — or evidence through pilot-testing (such as through think-alouds) are recommended to verify the presence of presentational confounds. Such testing would assess whether validated measures display reliability and validity with interspersion, labeling, and other formats.

Our findings point to a recommendation relating to the interactive effects of sample composition and respondent-centered scales. Particularly in terms of dimensionality, we found that interspersion leads to unsatisfactory levels of fit for non-student samples for respondent-centered scales, pointing to the downside of such a format. Clearly, special effort should be taken when surveying non-student adults to motivate respondents, simplifying the length and complexity of questionnaires, and reducing the use of abstractions. However, as noted above, this isolated result should be viewed with caution and subjected to further empirical testing.

So far, our discussion has focused on broad patterns that relate to cross-construct relationships and reflect differences in format conditions. However, narrower patterns may apply for individual scales and reflect differences in format conditions, and even narrower patterns may hold for items within individual scales and reflect differences in format conditions. As an example of the former, the ethnocentrism scale in the resequenced items condition led to a drop in fit in CFA models (Table 3). Similarly, the consumer independent judgment making subscale had a low test-retest correlation, and several subscales of material values displayed relatively low internal consistency under the contiguous paginated condition (Table 2), and perceived transaction value had low internal consistency in the interspersed condition (Table 9). These findings suggest caution in assuming that measures can be presented in ways that differ from their presentation at validation. As format factors involve using items and constructs in ways different from their presentation at validation, implications of our research extend to researchers involved in measure development and validation in efforts to address these issues, as well as to researchers who use these measures in conditions different from those employed for validation.

Finally, we note that our finding of the lack of impact of format has a limitation. In a strict scientific sense, it does not indicate that the impact is absent — rather, it indicates that the evidence does not indicate any impact. We have strived to include as many conditions and moderators as possible, along with relatively large sample sizes, across five separate studies. Further, we have examined a large number of statistics to detect impact. MGCFA has overwhelmingly suggested that measures are consistent. We have also constructed 95% confidence intervals to examine statistical differences. This examination clearly indicates a lack of impact of format on psychometric properties. Although this conclusion is based only on our studies, the evidence does suggest that implicit beliefs about impact of format require reexamination. Study 5 has limitations as well, as we studied only two conditions with a non-student sample. Moreover, the non-student sample was not representative of any larger population and was relatively small in size. Nevertheless, the study expands to non-student samples and emphasizes the importance of further research on this topic with such samples. In general, we note that all the studies would benefit from increased sample sizes for the statistical tests we employed.

In summary, researchers use a variety of different formats to structure questionnaires in theory testing. Such format choices reveal implicit beliefs that labeling, interspersion, pagination, sequencing, and/or pagination affects covariation between measures of constructs of interest. Our empirical investigation, spread across five studies, is, to our knowledge, the first comprehensive systematic examination whether such format choices matter, and if so, how. Our overwhelmingly consistent finding, based on our studies, is that format choices do not affect model fit, particularly for student samples, suggesting that researchers' implicit beliefs may need to be reexamined.

NOTE

1. Across the studies, we designed methods such that variations in administration did not create differences across conditions. However, we caution conservatively that a potential confound exists in this study between the questionnaire being completed at the middle versus end of administration and the conditions. As reported subsequently, the similar pattern of results here when compared to Study 1 suggests that being part of a study of longer duration did not affect the basic pattern.

REFERENCES

Bagozzi, R. P., &. Edwards, J. R. (1998). A general approach to representing constructs in organizational research. *Organizational Research Methods*, *1*(1), 45-87.

Bearden, W. O., Netemeyer, R. G., & Teel, J. E. (1989). Measurement of consumer susceptibility to interpersonal influence. *Journal of Consumer Research*, *15*, 473-481. doi:10.1086/209186

Bradlow, E. T., & Fitzsimons, G. J. (2001). Subscale distance and item clustering effect in self-administered survey: A new metric. *Journal of Marketing Research*, *38*(2), 254-261.

Burchell, B., & Marsh, C. (1992). The effect of questionnaire length on survey response. *Quality & Quantity*, *26*, 233-244. doi:10.1007/BF00172427

Byrne, B. M., Shavelson R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466. doi:10.1037/0033-2909.105.3.456

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306-307.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255. doi:10.1207/S15328007SEM0902_5

Churchill, G. A., Jr., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research*, *21*(4), 360-375.

Cote, J., & Buckley, M. R. (1988). Measurement error and theory testing in consumer research: An illustration of the importance of construct validation. *Journal of Consumer Research*, *14*(4), 579-582.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*(4), 407-422.

Dabholkar, P. A., Thorpe, D. I., & O Rentz, J. (1996). A measure of service quality for retail stores: Scale development and validation. *Journal of the Academy of Marketing Science*, *24*(1), 3-16.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman&Hall.

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

Gordon, M. E., L. Slade, A., & Schmitt, N. (1986). The "science of the sophomore" revisited: From conjecture to empiricism. *The Academy of Management Review*, *11*(1), 191-207.

Greenwald, A. G., & Banaji, M. R. (1989). The self as a memory system: Powerful, but ordinary. *Journal of Personality and Social Psychology*, *57*(1), 41-54.

Grewal, D., Monroe, K. B., & Krishnan, R. (1998). The effects of price-comparison advertising on buyers' perceptions of acquisition value, transaction value, and behavioral intention. *Journal of Marketing*, *62*(2), 46-59.

Groves, R. M. (1991). Measurement error across disciplines. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement error in surveys* (pp. 1-25). New York, NY: John Wiley.

Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*, 549-559. doi:10.1086/268687

Holbrook, M. B., & Batra, R. (1987). Assessing the role of consumer responses to advertising. *Journal of Consumer Research*, *14*, 404-420. doi:10.1086/209123

Kline, P. (2000). *The handbook of psychological testing*. New York, NY: Routledge.

Kline, T. J., Sulsky, L. M., &. Rever-Moriyama, S. D. (2000). Common method variance and specification errors: A practical approach to detection. *Journal of Psychology*, *134*(4), 401-421.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*, 312-320. doi:10.1037/0022-3514.55.2.312

Knowles, E. S., & 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.

Krishnamurthy, P., & Sujan, M. (1999). Retrospection versus anticipation: The role of the ad under retrospective and anticipatory self-referencing. *Journal of Consumer Research*, *26*(1), 55-69.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567. doi:10.1146/annurev.psych.50.1.537

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology*, *79*, 332-340. doi:10.1037/0021-9010.79.3.332

Lennox, R. D., & Dennis, M. L. (1994). Measurement error issues in substance abuse services research: Lessons from structural equation modeling and psychometric theory. *Evaluation and Program Planning*, *17*(4), 399-407.

Lichtenstein, D. R., Ridgway, N. M., & Netemeyer, R. G. (1993). Price perceptions and consumer shopping behavior: A field study. *Journal of Marketing Research*, *30*, 234-235. doi:10.2307/3172830

Little, T. D., Cunningham, W. A., Golan S., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*, 151-173. doi:10.1207/S15328007SEM0902_1

Malhotra, N. K., Kim, S. S., & Patil, A. (2006). Common method variance in IS research: A comparison of alternative approaches and a reanalysis of past research. *Management Science*, *52*, 1865-1883. doi:10.1287/mnsc.1060.0597

Manning, K. C., Bearden, W. O., & Madden, T. J. (1995). Consumer innovativeness and the adoption process. *Journal of Consumer Psychology*, *4*(4), 329-345.

McQuarrie, E. F., & Munson, J. M. (1986). The Zaichkowsky Personal Involvement Inventory: Modification and extension. In P. Anderson & M. Wallendorf (Eds.), *Advances in consumer research* (pp. 36-40). Provo, UT: Association of Consumer Research.

Mittal, B. (1995). A comparative analysis of four scales of consumer involvement. *Psychology and Marketing*, *12*, 663-682. doi:10.1002/mar.4220120708

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of Advertising*, *19*(3), 39-52.

Parameswaran, R., Barnett A., Greenberg, D., Bellenger, N., & Robertson, D. H. (1979). Measuring reliability: A comparison of alternative techniques. *Journal of Marketing Research*, *16*(1), 18-25.

Parasuraman, A., Zeithaml, V., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, *64*(1), 12-40.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*(3), 450-461.

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

Richins, M. L., & Dawson, S. (1992). A consumer values orientation for materialism and its measurement: Scale development and validation. *Journal of Consumer Research*, *19*(3), 303-316.

Shimp, T. A., & Sharma, S. (1987). Consumer ethnocentrism: Construction and validation of the CETSCALE. *Journal of Marketing Research*, *24*(3), 280-289.

Simmons, C. J., Bickart, B. A., & Lynch, J. G. (1993). Capturing and creating public opinion in survey research. *Journal of Consumer Research*, *20*(2), 316-329.

Smith, S. M., Haugtvedt, C. P., & Petty, R. E. (1994). Attitudes and recycling: Does the measurement of affect enhance behavioral prediction? *Psychology&Marketing*, *11*(4), 359-374.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78-90.

Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers*. San Francisco, CA: Jossey-Bass.

Szybillo, G. J., Binstok, S., & Buchanan, L. (1979). Measure validation of leisure time activities: Time budgets and psychographics. *Journal of Marketing Research*, *16*(1), 74-79.

Teas, R. K., & Agarwal, S. (2000). The effects of extrinsic product cues on consumers' perceptions of quality, sacrifice and value. *Journal of the Academy of Marketing Science*, *28*, 278-290. doi:10.1177/0092070300282008

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70.

Viswanathan, M. (2005). *Measurement error and research design*. Thousand Oaks, CA: Sage.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*, 956-972. doi:10.1177/0013164404268674

Zeithaml, V. A. (1988). Consumer perceptions of price, quality, and value: A means-end model and synthesis of evidence. *The Journal of Marketing*, *52*(3), 2-22.

APPENDIX A

Description of Experimental Conditions

Condition	Description
1. Contiguous	Items appear in the order in which they appeared in validity tests and related measures are contiguous
2. Contiguous and labeled	Same as condition 1, but measures are labeled
3. Contiguous and paginated	Same as condition 1, but measures are on separate pages
4. Interspersed	Items are completely interspersed in the questionnaire
5. Resequenced measures	Items are contiguous, but measures are resequenced
6. Resequenced and labeled measures	Items are contiguous, but measures are resequenced and labeled
7. Resequenced items, contiguous measures	Measures are contiguous, but items are resequenced
8. Resequenced items, contiguous, and labeled measures	Measures are contiguous and labeled, but items are resequenced

APPENDIX B

Models of Invariance
Six Models of Invariance in Order of Stringency of Requirements[a]

Model	Level of invariance	Description of invariance
1	Configural invariance	Similar pattern of factor loadings across conditions
2	Metric invariance	Same factor loadings across conditions
3	Scalar invariance	Intercepts are the same across conditions
4	Error variance invariance	Measurement error is the same across conditions
5	Factor variance invariance	Variance of latent constructs is equal across conditions
6	Factor covariance invariance	Covariance structure of latent constructs is the same across conditions

Note. [a] Testing for invariance requires the researcher to first compare the null model (i.e., a model without any constraints) against Model 1 (configural invariance) on the Bayesian information criterion, and then progressively compare Model 1 through to Model 6 on BIC to assess the type of invariance across conditions. Model 6 represents the most stringent level of invariance.

APPENDIX C

Bootstrapping Procedure

In order to test whether the reliability of measures is dependent on questionnaire formats, we bootstrapped the empirical data to form 95% confidence intervals around the estimated reliability coefficients. Bootstrapping is a commonly used statistical procedure to construct confidence intervals when the distribution of the statistic of interest is not known. Efron and Tibshirani (1993) provide a detailed review of bootstrapping procedures. We implemented a standard case re-sampling procedure, and constructed confidence intervals using percentile bootstrapping, as follows:

1. We first sampled, with replacement, *N* observations from the dataset with *N* observations. That is, if the dataset has 100 observations, we randomly drew 100 observations to construct a bootstrap sample. The "with replacement" procedure implies that we draw the first observation randomly, then put it back in the pool, pick the second observation, put it back in the pool, and so on until we sample 100 observations. This procedure may result in some observations being duplicated in a particular bootstrap sample while others may be omitted.
2. We then estimated the appropriate statistic using the bootstrap sample.
3. We repeated the above steps 500 to 1000 times to estimate each confidence interval. That is, we drew 500 to 1000 bootstrap samples from the original data. We repeated the procedure for 5000 resamples and achieved identical substantive results.
4. Finally, we calculated the percentiles for each of the statistics. In our case, we calculated the 2.5% and 97.5% percentiles to give us 95% confidence intervals around our estimates.

We then examined the confidence intervals of the estimated statistics under the conditions of each study. All bootstrapping results are available from the authors.