Article

Educational and Psychological Measurement
2016, Vol. 76(6) 912–932

DOI: 10.1177/0013164415627349
epm.sagepub.com

Survey Satisficing Inflates Reliability and Validity Measures: An Experimental Comparison of College and Amazon Mechanical Turk Samples

Tyler Hamby[1] and Wyn Taylor[1]

Abstract

This study examined the predictors and psychometric outcomes of survey satisficing, wherein respondents provide quick, "good enough" answers (satisficing) rather than carefully considered answers (optimizing). We administered surveys to university students and respondents—half of whom held college degrees—from a for-pay survey website, and we used an experimental method to randomly assign the participants to survey formats, which presumably differed in task difficulty. Based on satisficing theory, we predicted that ability, motivation, and task difficulty would predict satisficing behavior and that satisficing would artificially inflate internal consistency reliability and both convergent and discriminant validity correlations. Indeed, results indicated effects for task difficulty and motivation in predicting survey satisficing, and satisficing in the first part of the study was associated with improved internal consistency reliability and convergent validity but also worse discriminant validity in the second part of the study. Implications for research designs and improvements are discussed.

Keywords

satisficing, MTurk, survey design, psychometrics

Since Campbell and Fiske's (1959) influential article, social scientists have acknowledged the importance of convergent and discriminant validity in questionnaires. In

[1]University of Texas at Arlington, TX, USA

Corresponding Author:
Tyler Hamby, Department of Psychology, University of Texas at Arlington, Arlington, TX 76019, USA.
Email: tylernhamby@gmail.com

short, a self-report scale should correlate strongly with other measures of the same construct (convergent validity) but should correlate weakly with unrelated constructs (discriminant validity). For example, one would expect different extraversion questionnaires to be strongly correlated, but extraversion should not be strongly correlated (in absolute value) with a neuroticism questionnaire. A scale may be shown to have a high coefficient alpha, a measure of internal consistency reliability, and to correlate strongly with related or synonymous constructs; however, if it also correlates strongly with conceptually independent constructs, it would lack discriminant validity. Results from such a survey should obviously be questioned. In the present research, we aim to investigate some of the attributes of scales and samples that increase internal consistency reliability and convergent validity at the expense of discriminant validity.

Survey Satisficing

To identify the causal, psychological determinants of these psychometric outcomes, researchers have employed methods and theory from cognitive psychology to develop a useful framework for conceptualizing the process of responding to survey items. Borrowing from Simon's (1956) theory of satisficing—which described the tendency to seek quick, ''good enough'' answers rather than invest time and/or resources searching for the best or optimal answer—Krosnick (1991) distinguished between *optimizing* and *satisficing* approaches to performance on questionnaires. To answer survey questions optimally, or to optimize, people presumably go through four steps (Schwarz & Strack, 1985; Tourangeau & Rasinski, 1988): They must interpret the question and deduce its intent, search their memories for relevant information, integrate the information into a single judgment, and then select the most appropriate survey response-scale category for that judgment. Ideally, subjects optimize and thoroughly complete each step for each survey item presented.

In reality, respondents often satisfice, or settle for a ''good enough'' answer, by skipping or expending minimal effort on these cognitively demanding steps. The likelihood of satisficing is thought to increase with (1) decreased respondents' abilities, (2) decreased respondents' motivation, and (3) increased task difficulty (Krosnick, 1991, 1999). Samples prone to satisficing behavior may be less educated, less intelligent, or simply lack motivation. Conditions conducive to survey satisficing include design characteristics that increase the cognitive difficulty of responding to survey items. Thus, the theory of satisficing provides for testable hypotheses.

Not surprisingly, survey-satisficing behaviors are assumed to result in less valid survey data. For example, satisficing respondents may repeatedly select the same response-scale category for items that they superficially assess as measuring the same construct rather than carefully thinking about and responding to each item (Krosnick, 1991; Krosnick & Alwin, 1988). This seemingly rational, yet problematic, shortcut of not differentiating between similar items has been operationalized as survey satisficing (Lelkes, Krosnick, Marx, Judd, & Park, 2012). Of course, if the respondents select the same option for each item in a scale, they are increasing the scale's internal

consistency reliability in that sample, but the satisficing participants' scores on that scale are less meaningful (i.e., valid) than if they had considered each item separately. Though studies have reported many forms of survey-satisficing behavior—for example, excessively selecting the middle alternative (Narayan & Krosnick, 1996) and ''speeding'' (Zhang & Conrad, 2013)—the current research examined ''nondifferentiation'' between similar items as the satisficing measure.

Research on Samples of Convenience

People have natural decision-making tendencies—to either satisfice or optimize—when faced with everyday, real-life problems, each approach having its own set of advantages and disadvantages (Simon, 1956). Even so, research design and the context in which a respondent completes a survey can aggravate natural ''satisficers'' (i.e., those respondents predisposed to satisfice on surveys) even if their intent is to pay careful attention and can trigger satisficing behavior in people generally inclined to optimize. Historically, behavioral research has relied on college students as samples of convenience that, while useful and informative, do not necessarily generalize to the larger population. For instance, in a second-order meta-analysis (i.e., a meta-analysis of meta-analyses), Peterson (2001) found the behavioral research responses of college students to be more homogenous and to differ both in effect sizes and sometimes direction, compared with community, adult samples. College students, by their nature, have to at least be motivated enough to get into a university; therefore, one would expect they would be less likely to satisfice in the first place and that they may be more resistant to contextual and design characteristics known to trigger satisficing behaviors, regardless of their trait-level tendencies.

Online survey sites are a recent development and boon, especially for survey research; but data collected from these sites should also be considered samples of convenience. While the differences between college and community samples have been well studied, the specific strengths and weaknesses of for-pay, online survey samples from sites like Amazon Mechanical Turk (MTurk; see www.MTurk.com) require further investigation to ensure proper interpretation and consideration of larger implications. Many researchers have recommended using MTurk samples even while recognizing generalizability issues (Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013; Crump, McDonnell, & Gureckis, 2013; Goodman, Cryder, & Cheema, 2012; Paolacci & Chandler, 2014; Paolacci, Chandler, & Ipeirotis, 2010; Rand, 2012; Rouse, 2015).

However, a standout problem with for-pay survey sites, related to survey satisficing, is the effect of these sites' financial incentives on respondents' behavior. In a general meta-analysis, Cerasoli, Nicklin, and Ford (2014) showed that extrinsic (such as financial) incentives decrease performance *quality* and increase performance *quantity*. Extending this finding to for-pay survey sites, the best way to make the most money is to complete as many surveys as quickly as possible (Malhotra, 2008; Zhang & Conrad, 2013). This fundamentally discourages desired behaviors—such as careful

examination of survey items (Tourangeau & Rasinski, 1988)—and incentivizes undesirable survey-satisficing behavior. Indeed, in a large, multinational study (Barge & Gehlbach, 2012), financial incentives increased survey satisficing (nondifferentiation and speed) and survey completion (more items completed), while undermining data quality. Moreover, many other researchers have also suggested financial incentives increase survey completion rates and times, meaning respondents—who might otherwise have quit—are retained and speed through a survey just to finish it (e.g., Barge & Gehlbach, 2012; Buhrmester et al., 2011; Casler et al., 2013; Crump et al., 2013). Consequentially, it may be that financial incentives act to retain a group of satisficers, who might otherwise have been removed from the sample as incomplete data points.

The financial motivation offered by MTurk, and other online survey sites, may predispose these samples to satisficing behavior; in comparison, college students samples should be more likely to optimize due to the personality and motivational characteristics mentioned above. However, in order to attribute any differences between these samples to personality or motivation, it is necessary to control for demographic variables because research has shown that MTurk samples also differ from college samples in this regard (e.g., Paolacci & Chandler, 2014). Still, the differing characteristics of these samples of convenience provide testable predictions related to survey satisficing.

Research on Survey Satisficing

Some research studies have confirmed Krosnick's (1991) theory that three factors—respondent ability, respondent motivation, and task difficulty—do tend to predict satisficing behavior. An experimental study (Krosnick, Narayan, & Smith, 1996) and a meta-analytic study (Narayan & Krosnick, 1996) have correlated measures of intelligence with various survey-satisficing behaviors. More specific to our purposes, other studies (e.g., Krosnick & Alwin, 1988; Zhang & Conrad, 2013) have shown that less educated respondents were more likely to repeatedly select the same response-scale categories for items that measure similar constructs.

Oppenheimer, Meyvis, and Davidenko (2009; see also Krosnick et al., 1996) showed that satisficing behavior was associated with subjects' motivation to think about the items, as measured by Need for Cognition (NFC; Cacioppo, Petty, & Kao, 1984). Interestingly, Lelkes et al. (2012) showed that anonymous (and presumably less motivated) participants were more likely to exhibit nondifferentiation in responses to survey items but only toward the end of the survey. People become fatigued by the end of a survey, and a significant difference was only found when the less motivated group became fatigued and therefore especially unmotivated. This study raised the possibility that the vulnerabilities for satisficing behavior tend to interact to predict satisficing in a particular setting. Similarly, another study (Krosnick et al., 1996) showed that various measures of motivation predicted nondifferentiation in both main effects and interaction effects.

Fewer studies have investigated a link between survey-satisficing behavior and task difficulty, but some researchers have suggested that certain response-scale formats may be more difficult for participants. In particular, response scales that only have the endpoints labeled may be more difficult for respondents to use than fully labeled response scales (Krosnick & Presser, 2010). They also suggest that it may be more challenging for participants to choose the optimal response-scale category for items with lengthier response scales (e.g., 7-point scales as opposed to 4- or 5-point scales). These scale formats may then encourage satisficing behaviors. Additionally, whether or not the above survey-response tasks *feel* difficult to a respondent may be a function of their cognitive capabilities (Hamby, 2015). So, intelligence (measured as education level) may interact with task difficulty to predict satisficing behavior. In general, the above-mentioned literature review suggests that the various vulnerabilities to survey satisficing may interact to predict satisficing behavior.

Compared with the growing collection of studies that document the causes of satisficing, less is known about how survey satisficing affects a study's results. In particular, we know of no study to examine the correspondence between survey satisficing and reliability, convergent validity, and discriminant validity. This gap in the research literature on survey satisficing motivated the present study.

Present Study

The purpose of this study is twofold: (1) to first correlate predictors of survey satisficing with nondifferentiation and (2) to then examine the psychometric consequences of nondifferentiation and the predictors of satisficing. For this purpose, we administered a personality survey to participants who were randomly assigned to survey conditions that varied in response-scale length (i.e., the number of response-scale categories) and label format, and they were either given course credit or a small amount of money for their participation. The survey had two distinct parts: In the first, we determined whether the respondent evidenced satisficing behavior, operationalized as nondifferentiation between similar items; in the second, we quantified the effects of that behavior on the interitem correlations for subsequent questionnaires. In particular, we examined correlations between items of the same questionnaire, items from different questionnaires that purport to measure the same construct, and items from questionnaires that measure different constructs because these correlations are determinants of internal consistency reliability, convergent validity, and discriminant validity, respectively.

Hypotheses

Based on the aforementioned review, we predicted that being less motivated—operationalized as being paid a small amount of money to complete the survey via MTurk—and ability—assumed to be associated with less education—would both predict nondifferentiation. We predicted that the task difficulty of having response-

scales with only endpoints labeled or with too many categories would predict satisficing behavior as measured by nondifferentiation. We additionally hypothesized that these factors would interact to predict satisficing behavior. For the second part of our study, we predicted that having demonstrated satisficing behavior (measured as nondifferentiation) in the first part of the study would be associated with improved internal consistency reliability and convergent validity but impaired (i.e., artificially inflated) discriminant validity. We also hypothesized that the aforementioned vulnerabilities to satisficing—low cognitive ability, low motivation, and task difficulty—would each positively predict the interitem correlation sizes in main effects and in interactions with prior nondifferentiation. Last, we predicted that these effects would be found even after controlling for demographic differences between MTurk and student samples.

Method

Participants

The participants in this study were 893 U.S. workers on MTurk who took our surveys in exchange for financial compensation ($ 0.50) and 479 undergraduate students at a large, southern university who took our surveys in exchange for course credit. After screening the data for outliers and excessive missing values, we had 882 MTurk respondents (483 females, 397 males, 2 no response) and 465 university students (333 females, 131 males, 1 no response). The MTurk sample had a more even gender distribution (55% female) than the university sample (72% female), $\chi^2(1) = 36.20$, $p < .001$, $\varphi = .16$, and they were older ($M = 35.30$, $SD = 12.30$) than the university sample ($M = 20.58$, $SD = 3.77$), $t(1, 150.70) = 32.71$, $p < .001$, Cohen's $\bar{d} = 1.62$. However, the university sample was somewhat more ethnically diverse: The MTurk sample had 76% White participants, whereas the university sample had only 34% White participants, $\chi^2(1) = 228.17$, $p < .001$, $\varphi = .41$. The MTurk sample had varying education levels: 11 (1%) had not completed high school, 90 (10%) were high school graduates, 343 (39%) had completed some college, 326 (37%) had bachelor's degrees, 98 (11%) had master's degrees, and 13 (2%) had doctoral degrees. So, 437 participants (50%) held college degrees.

Materials and Procedures

Participants signed up for the study on the Amazon Mechanical Turk website or through the psychology department's student subject pool, where they clicked a link to a webpage that we set up. The webpage had the informed consent, and a ''Start Survey'' button that randomly sent the participant to one of six survey conditions. All participants took two sets of questionnaires. The first set of questionnaires contained seven scales, totaling 59 items with 13 (22%) reverse-scored. Every participant was administered the 8-item extraversion and 8-item neuroticism subscales of the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991), the 13-item Revised

Self-Monitoring Scale (RSMS; Lennox & Wolfe, 1984), the 10-item Rosenberg's Self-Esteem Scale (SES; Rosenberg, 1965), the 5-item Satisfaction with Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985), and the 10-item dysphoria and 5-item social anxiety scales of the Inventory of Depression and Anxiety Symptoms (IDAS; Watson et al., 2007). We used the first set of questionnaires to examine predictors of satisficing behavior as operationalized by nondifferentiation: specifically whether or not a participant selected the same response-scale category for all items in at least one of the seven scales.

For the first set of questionnaires, respondents were randomly assigned to one of six conditions. They either had all response-scale categories labeled (ALL; $N = 699$; 52%) or only the endpoints labeled (END; $N = 648$; 48%), and they had all questions with four ($N = 458$; 34%), five ($N = 439$; 33%), or seven ($N = 450$; 33%) response-scale categories. We decided on these particular response-scale lengths because few authors use more than seven categories. For instance, in a recent meta-analysis encompassing research from 24 psychology, marketing, management, and education journals, only 31 (1%) of 2,524 alpha coefficients arose from scales with eight or more response-scale categories, and most (91%) used either five or seven categories (Peterson & Kim, 2013).

The second set of questionnaires consisted of 30 items (50% reverse-scored): 15 extraversion and 15 neuroticism items from the International Personality Item Pool (IPIP; 2001). This section only had two conditions: The response-scale label format matched that which was assigned for the first set of questionnaires (ALL or END). However, for both conditions, five items had four response-scale categories, five items had five response-scale categories, and five items had seven response-scale categories for both extraversion and neuroticism. We used all three response-scale lengths to counteract any bias that the response-scale length condition for the first set of questionnaires may have on the responses to the second set of questionnaires. We used these questionnaires to quantify the impact that satisficing behavior and other predictors had on internal consistency reliability, convergent validity, and discriminant validity by examining interitem correlations between items within a particular scale (e.g., two of the extraversion items with four response categories), between items in two related scales (e.g., one extraversion item with four response categories and one extraversion item with five response categories), and between items of two unrelated scales (e.g., one extraversion item and one neuroticism item), respectively.

In choosing the response-scale category labels, we attempted to divide the response-scale continuum into equidistant categories (Hamby & Levine, 2016). So, we selected labels based on past research that has mapped the psychological location of category descriptors onto a response-scale continuum (Bass, Cascio, & O'Conner, 1974; Dobson & Mothersill, 1979). For the BFI scales, RSMS, SES, and SWLS, we used 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Neither agree nor disagree*, 5 = *Slightly agree*, 6 = *Agree*, and 7 = *Strongly agree* for the 7-point scale. For the 5-point scale, we omitted response-scale options 3 = *Slightly disagree* and 5 = *Slightly agree*, and for the 4-point scale, we removed the middle category.

For the IDAS scales, we used 1 = *Not at all*, 2 = *A little bit*, 3 = *Somewhat*, 4 = *Moderately*, 5 = *Quite a bit*, 6 = *Very*, and 7 = *Extremely* for the 7-point scale. For the 5-point scale, we omitted 3 = *Somewhat* and 6 = *Very*, and for the 4-point scale, we removed 4 = *Moderately*. Finally, for the IPIP scales we used 1 = *Very inaccurate*, 2 = *Inaccurate*, 3 = *Moderately accurate*, 4 = *Neither accurate nor inaccurate*, 5 = *Moderately accurate*, 6 = *Accurate*, 7 = *Very accurate*. For the 5-point scale, we omitted 2 = *Inaccurate* and 6 = *Accurate*, and for the 4-point scale, we omitted the middle category.

Results

Predictors of Satisficing Behavior in First Set of Questionnaires

We used a sequential logistic regression analysis to examine how well our predictor variables—sample type, response-scale label format, and number of response-scale categories—and our demographic covariates—gender, age, and ethnicity (Asian, Black, White, Latino, or other)—predicted our measure of satisficing behavior: whether or not respondents gave identical responses to one or more scales of the first set of seven scales.[1] A minority, $N = 485$ (36%), did give uniform responses to at least one scale. First, using the likelihood ratio test, we determined that the model with only the predictors, $\chi^2(5, N = 1{,}339) = 52.77$, $p < .001$, and the model with only the covariates, $\chi^2(6, N = 1{,}339) = 24.78$, $p < .001$, each improved on the intercept-only model. Next, we found that the model with main effects for both predictor variables and covariates improved on the model with only covariates, $\chi^2(5, N = 1{,}339) = 34.63$, $p < .001$, but as predicted, it did not improve on the model with only predictors, $\chi^2(6, N = 1{,}339) = 6.64$, $p = .36$. Hence, we excluded the demographic covariates from the model. Last, contrary to hypothesis, the model with predictor variables and all two-way interactions did not significantly improve on the main effects model, $\chi^2(8, N = 1{,}339) = 3.54$, $p = .90$. Thus, we examine only the main effects.

Table 1 summarizes the results of the logistic regression. In line with hypothesis, the sample type did predict satisficing behavior, $p < .001$. As expected, compared with the university sample, the MTurk samples with, $p < .001$, and without college degrees, $p < .001$, were each more likely to evidence satisficing behavior, but counter to prediction, the non-college-educated MTurk sample was not more likely to satisfice than the college-educated MTurk sample, $p = .68$. This supports our hypothesis that the financially motivated MTurk samples were more prone to satisficing behavior than the university sample, but it does not support the hypothesis that intelligence would predict satisficing.

However, unexpectedly, label format did not significantly predict satisficing behavior, $p = .37$, and also unexpectedly, response-scale length was negatively, rather than positively, associated with nondifferentiation, $p < .001$. The respondents in the conditions with five categories, $p < .01$, and seven categories, $p < .001$, were less likely to give identical responses to at least one scale than those in the four-category condition. Thus, task difficulty, neither in terms of having more response-scale categories nor

Table 1. Logistic Regressions for Sample-Type, Response-Scale Label Format, and Number of Response Categories Predicting Satisficing Behavior.

Predictors	Wald χ^2	*df*	*p*	Levels of predictor	*b*	*p*	OR
Sample	25.673	2	.000	MTurk—No degree	−.057	.684	0.945
				MTurk—Degree	.668	.000	1.950
				University sample	−.611	.000	0.543
Label	0.82	1	.365	ALL	.105	.365	1.111
No. of categories	26.977	2	.000	Four	.364	.009	1.439
				Five	.378	.010	1.459
				Seven	−.742	.000	0.476

Note. All levels of the predictor variables are compared with the level directly below it in the list for that particular predictor; those at the bottom of the list (University Sample and Seven) are compared with those at the top of the list (MTurk—No Degree and Four). Sample = either MTurk with no college degree, MTurk with college degree, or university student; Label = whether the scale had every point labeled or only endpoints labeled; No. of categories = either four, five, or seven response-scale categories; OR = odds ratio.

END scales, predicted satisficing behavior as expected. The negative relationship between response-scale length and satisficing behavior may simply be an artifact of our definition of satisficing. For example, the probability of randomly selecting the same response-scale category two out of two times is higher when there are four (1/16), rather than seven categories (1/49).

Correlations for Second Set of Questionnaires

Table 2 presents the correlation matrices for the five-item IPIP extraversion and neuroticism scales with four, five, and seven response-scale categories, and these matrices are separated by sample type (MTurk with or without college degrees or the university student sample). The results clearly show that the reliabilities (extraversion $M = .80$; neuroticism $M = .77$) and convergent validity correlations between the different measures of extraversion ($M = .83$) and neuroticism ($M = .78$) are relatively high for five-item scales. More important, the correlations between the extraversion and neuroticism measures were extremely high ($M = -.42$), demonstrating poor discriminant validity. In particular, the discriminant validity correlations (i.e., correlations between extraversion and neuroticism scales) were affected by sample type: The MTurk sample without college degrees ($M = -.48$) and with college degrees ($M = -.43$) generally had stronger correlations and, thus, worse discriminant validities than the student sample ($M = -.33$).

Psychometric Consequences of Satisficing in Second Set of Questionnaires

We next examined how prior satisficing behavior (i.e., nondifferentiation), sample-type, and label format influenced the reliabilities and validity correlations presented

Table 2. Correlations and Reliabilities for International Personality Item Pool Extraversion and Neuroticism Measures With Four, Five, and Seven Response Categories Separated by Sample Type.

	MTurk—No degree						MTurk—Degree						University					
Scale	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1. 4-Ex	.84						.80						.78					
2. 4-Ne	−.57	.82					−.53	.78					−.49	.68				
3. 5-Ex	.84	−.48	.82				.81	−.42	.78				.76	−.33	.73			
4. 5-Ne	−.54	.85	−.45	.79			−.48	.80	−.41	.77			−.42	.75	−.36	.67		
5. 7-Ex	.86	−.52	.85	−.50	.80		.82	−.46	.83	−.44	.80		.78	−.35	.80	−.32	.73	
6. 7-Ne	−.47	.79	−.34	.83	−.42	.84	−.42	.76	−.34	.81	−.39	.82	−.28	.63	−.16	.69	−.20	.71

Note. The statistics on the diagonals are alpha reliabilities. All other statistics are correlations between scales. MTurk = Amazon Mechanical Turk; No degree = no college degree; Degree = college degree; University = university student sample; 4 = four response-scale categories; 5 = five response-scale categories; 7 = seven response-scale categories; Ex = extraversion; Ne = neuroticism.

Table 3. Repeated-Measures ANOVAs for Satisficing Behavior, Sample Types, and Label Format Predicting Interitem Correlations Between Same or Different Constructs From the International Personality Item Pool Extraversion and Neuroticism Scales.

	Same construct			Different construct		
Predictors	Wilks's Λ	p	η^2	Wilks's Λ	p	η^2
Satisficing	.129	.000	.871	.071	.000	.929
Sample	.262	.000	.738	.343	.000	.657
Label	.910	.000	.090	.397	.000	.603
Satisficing × Sample	.678	.000	.322	.974	.055	.026
Satisficing × Label	.809	.000	.191	.842	.000	.158
Sample × Label	.529	.000	.471	.868	.000	.132
Satisficing × Sample × Label	.683	.000	.317	.700	.000	.300

Note. Same construct = both items measure extraversion or both items measure neuroticism; Different construct = one item measures extraversion and one item measures neuroticism; Satisficing = whether or not the participant gave identical responses to one or more scale in the first set of questionnaires; Sample = either MTurk with no college degree, MTurk with college degree, or university student; Label = whether the scale had every point labeled or only endpoints labeled; × = indicates an interaction effect; ANOVA = analysis of variance.

in Table 2. For this purpose, we conducted two sets of 2 (satisficing behavior: zero questionnaires with uniform responses, one or more questionnaire with uniform responses) × 3 (sample-type: MTurk with no degree, MTurk with college degree, university undergraduate student sample) × 2 (label format: ALL, END) repeated-measures analyses of variance (ANOVAs) with the 435 pairs of interitem correlations as the dependent variables.[2] One ANOVA, which examined internal consistency reliability and convergent validity, consisted of all 210 correlations between items that measure the same construct (either extraversion or neuroticism), and the other ANOVA, which examined discriminant validity, used the 225 correlations between items that measure different constructs (one measured extraversion and the other measured neuroticism). We included all main effects and all interactions; the results are shown in Table 3.

Table 3 shows that all main effects and interactions were statistically significant, $p < .001$, except for the interaction between sample-type and prior satisficing behavior in the ANOVA for items measuring different constructs, $p = .06$. As predicted, those conditions assumed to be associated with satisficing behavior—giving uniform responses to at least one of the first seven questionnaires, being sampled from MTurk, and END label format—tended to have stronger correlations between same *and* different constructs. Hence, satisficing behavior appears to be associated with improved internal consistency reliability and construct validity but also worsened discriminant validity.

These effects seem to actually be stronger for the correlations between items of different constructs; the average correlations for the 12 conditions ranged from −.08

Table 4. Means and Post-Hoc Comparisons of Repeated-Measures ANOVAs for Satisficing Behavior, Sample Type, and Label Format Predicting Interitem Correlations Between Same or Different Constructs From the International Personality Item Pool Extraversion and Neuroticism Scales.

	Same construct				Different construct			
	ALL		END		ALL		END	
Predictors	None	Sat	None	Sat	None	Sat	None	Sat
MTurk—No degree	.41	.55	.41	.57	−.16	−.34	–.24	−.39
MTurk—Degree	.37	.56	.35	.49	−.15	−.29	−.15	−.37
University	.31[a]	.32[a]	.32	.48	−.08	−.19	−.12	−.33

Note. All pairwise comparisons of the effect of satisficing behavior are significant at $p < .001$ except those marked with superscript "a". Same construct = both items measure extraversion or both items measure neuroticism; Different construct = one item measures extraversion and one item measures neuroticism; ALL = fully labeled response scale; END = endpoint labeled response scale; None = no prior satisficing behavior; Sat = prior satisficing behavior; MTurk = Amazon Mechanical Turk; No degree = no college degree; Degree = college degree; University = university student sample; ANOVA = analysis of variance.
[a]$p = .44$.

to −.39 for items measuring different constructs and from .31 to .57 for items of the same construct. In particular, using END scales increased correlation size more for pairs of items measuring different constructs (discriminant validity correlations; END $M = -.27$; ALL $M = -.20$) than for items measuring the same construct (reliability and convergent validity correlations; END $M = .44$; ALL $M = .42$). As measured by partial-eta size, prior satisficing behavior was the strongest predictor of reliability and convergent validity correlations (satisficers $M = .49$; nonsatisficers $M = .36$) and of discriminant validity correlations (satisficers $M = -.32$; nonsatisficers $M = -.15$). Last, using Bonferroni adjusted post-hoc comparisons, the three sample types all differed, at $p < .001$, for reliability and convergent validity correlations (MTurk-no degree $M = .48$; MTurk-degree $M = .44$; university $M = .35$) and for discriminant validity correlations (MTurk–No degree $M = -.28$; MTurk–Degree $M = -.24$; university $M = -.18$). These results strongly support the hypotheses that our measures of previous satisficing behavior, task difficulty, lower levels of intelligence, and lowered motivation were all predictive of inflated interitem correlation sizes for measures of both the same and different constructs.

Table 4 provides the means for all levels of the predictors and the Bonferroni adjusted post-hoc comparisons—between those who were satisficers and those who were not—for the three-way interactions at each combination of the other two variables: sample-type and response-scale label format. It is interesting to note that the only nonsignificant effect was for those assumed to be least likely to evidence satisficing behavior: the university students in the ALL condition. In general, the interactions revealed that those conditions associated with satisficing behavior generally had a multiplicative, rather than additive, effect. Thus, confirming our hypothesis,

predictors of satisficing behavior are apparently especially impactful for those who are already prone to satisfice.

Discussion

The current research first examined how well ability, motivation, and task difficulty predicted satisficing behavior—as evidenced by respondents' nondifferentiation (giving similar ratings to superficially similar survey items) for a set of six questionnaires. Then, it examined the impact of satisficing behavior on those initial questionnaires, motivation, task difficulty, and ability on interitem correlations between items of the same scale (internal consistency reliability), of similar scales (convergent validity), and of different scales (discriminant validity) for a second set of two questionnaires.

Predictors of Satisficing Behavior

Participant Ability and Motivation. The results did support our hypothesis that the financially motivated MTurk samples would be more prone to satisficing behavior than the university sample. The best way to make the most money taking MTurk surveys is to complete as many as possible, as quickly as possible. Indeed, survey ''speeding'' (i.e., giving answers very quickly) and item nondifferentiation are highly correlated indicators of survey satisficing and low response quality (Malhotra, 2008; Zhang & Conrad, 2013). This inherent, compelling motivation to satisfice must be taken into consideration by researchers when using paid internet samples.

Counter to prediction, the college educated MTurk sample was no less likely to satisfice than the MTurk sample without college degrees. Assuming that education level is a valid measure of ability, this finding seems to suggest that respondent ability and survey satisficing are independent, which contradicts prior research (e.g., Krosnick & Alwin, 1988; Zhang & Conrad, 2013). However, lower education in the MTurk sample may not serve as an indicator of lower intelligence. As a survey site that offers financial rewards for participation, MTurk clearly attracts participants with free access to costly technology and who are computer savvy; but more important, this sample consists of people confident enough in their own ability they expose themselves repeatedly to surprise survey questions and tasks. It is entirely possible that the intelligence range in this sample is simply limited. Alternatively, these results could be interpreted as suggesting that financial motivation to satisfice may indeed trump individual differences and tendencies; that is, participants' cognitive ability and their decision strategy tendencies. This is a question for further research.

Task Difficulty. Response-scale label format (END or ALL) and having more response-scale categories were each assumed to be associated with task difficulty and therefore satisficing behavior (Krosnick & Presser, 2010). However, response-scale label format was not associated with satisficing in this study, and although the number of response-scale categories did predict satisficing behavior, the relationship was opposite of expectation. That is, having more response-scale categories from which

to choose predicted *less* satisficing behavior. As described above, this unanticipated result may be a consequence of our operational definition of satisficing (i.e., nondifferentiation), and perhaps an effect would have been found with a different measure for satisficing behavior. Future research should determine whether response-scale length predicts different measures of satisficing behavior.

Reliability and Validity

Prior Satisficing Behavior. We next examined the impact of prior satisficing behavior and the predictors of prior satisficing behavior on internal consistency reliability, convergent validity, and discriminant validity by analyzing inter-item correlations between two sets of questionnaires. As we expected, prior satisficing behavior (nondifferentiation) predicted subsequently higher reliabilities, as well as convergent *and* discriminant validities. In terms of effect size, prior satisficing behavior was actually the strongest predictor of both convergent and discriminant validity. This pattern has the surprising, and unfortunate, implication that high-scale reliabilities may sometimes signal that the scales have poor validity. The idea that scale developers may improve reliability at the expense of validity is nothing new (Nunnally & Bernstein, 1994); instead, we suggest that characteristics of the sample (e.g., intelligence and how the sample was collected) and the survey (e.g., response-scale label format) may cause or exacerbate these phenomena by means of increasing survey-satisficing behavior.

Participant Ability and Motivation. Next, confirming hypotheses regarding ability and motivation, sample-type—MTurk participants either with or without degrees or university students—did predict both convergent and discriminant validity, such that the less educated and less motivated groups tended to have stronger correlations in general. This finding has the clear implication that, while using MTurk will provide a quick and cheap sample, the quality of the resulting data may suffer.

Task Difficulty. Considering task difficulty, the number of response-scale categories did not appear to directly impact internal consistency reliability, convergent validity, or discriminant validity, but correlations were stronger for scales with END label formats in general. In particular, discriminant validity correlations between extraversion and neuroticism scales—two constructs that should *not* be strongly correlated—were particularly strong for END scales. This finding is consistent with Krosnick's (1999) assertion that endpoint labeled response scales are more difficult for respondents.

Interactions Between Predictors of Satisficing. Additionally, we examined the interactions between the aforementioned predictors—sample type and label format—and prior satisficing behavior with post-hoc comparisons. In general, these predictors interacted with satisficing behavior such that having shown signs of prior satisficing behavior was particularly predictive of future satisficing behavior for those conditions presumed to be affiliated with satisficing behavior (i.e., END label formats and MTurk participants). Though these results tell a complicated story about the

predictors of survey satisficing, they do dovetail nicely with prior research. Studies have shown that inherent vulnerabilities to satisfice interacts with extrinsic conditions conducive to satisficing behavior, such as fatigue (Lelkes et al., 2012) and certain linguistic characteristics of the items (Krosnick et al., 1996). The results are also comparable to Zhang and Conrad's (2013) finding that less educated respondents who engaged in prior satisficing behavior—operationally defined as answering survey items very quickly—were particularly likely to engage in further satisficing behavior (nondifferentiation).

To summarize, conditions assumed to be associated with satisficing behavior—giving uniform responses to at least one of the first seven questionnaires (nondifferentiation), being less educated (ability), being sampled from MTurk (low motivation), and having an END response-scale label format (task difficulty)—did indeed have stronger correlations between same (convergent validity) *and* different (discriminant validity) constructs, and these conditions apparently have an interactive effect. Hence, satisficing behavior appears to be associated with improved internal consistency reliability and convergent validity but also *worsened* discriminant validity.

Implications

For-Pay Survey Sites. Although extant research tends to be favorable for MTurk samples (e.g., Casler et al., 2013; Crump et al., 2013; Rouse, 2015), the findings from the present research have troubling implications for internet survey sites because they incentivize people to complete the as many surveys in as little time as possible; clearly this precludes careful examination of survey items and contemplation of responses. If financial motivation is sufficient to invite satisficing behavior, as was found in this research, these for-pay survey sites will have to be seen for what they are: sources for quick, inexpensive samples that may not necessarily generalize to other populations.

Attention differences. Prior research has shown that MTurk samples were less likely to pay attention to experimental instructions (Oppenheimer et al., 2009) and were more likely to cheat (Goodman et al., 2012). But measuring attention levels and screening out participants who do not pay attention may avoid these problems to some degree (Peer, Vosgerau, & Acquisti, 2014). Perhaps these efforts would have prevented the differences in interitem correlation size between the student and the MTurk sample in the present study.

Individual differences. As stated above, unlike prior research (Krosnick & Alwin, 1988; Zhang & Conrad, 2013), education level did not predict nondifferentiation, which may be a result of a range restriction of intelligence in MTurk samples. However, since the sort of people who take surveys for-pay have been shown to be relatively homogenous on some personality traits (e.g., self-esteem and extraversion; Goodman et al., 2012), it is plausible that MTurk participants may be more prone to be high in intelligence, though this hypothesis requires empirical support. If shown to be true, the decreased range of intelligence could be another factor that limits the generalizability of any research that utilizes paid survey respondents.

Survey Design. Survey designs can have a huge impact on whether or not respondents will satisfice or optimize their answers; high task difficulty is known to hamper the attentiveness of participants, especially those with low cognitive ability and/or motivation (Krosnick, 1991). While it is well known that high reliability is not necessarily indicative of desirable validity, this research suggests that poor scale validity could potentially be improved by study design changes that decrease satisficing behavior.

The present research presents a strong case against using END labeled scales; when using END scales, interitem correlations may be inflated, particularly for scales that measure dramatically different constructs. Labeling all the numbers in a rating scale—*strongly agree, agree, neither agree nor disagree, disagree, strongly disagree*—versus labeling only the ends, appears to require less thought on the part of respondents and allow for more nuanced responses. We agree with Krosnick (1999) that fully labeled scales have superior validities, and this seems to be especially true for respondents already prone to satisfice.

An important implication of this study is that simplifying a survey may produce more valid results for a more general population. For instance, low ability and/or unmotivated samples should not be administered long or difficult questionnaires, and they should not be given lengthy or complicated survey items (Hamby & Ickes, 2015). Additionally, attention checks appear to be a discreet design element that can detect satisficing behavior and facilitate sample filtering if necessary; they can also increase task attention, especially in participants most likely to satisfice (Peer et al., 2014). Efforts to simplify research designs and the addition of attention checks to help focus participants on a survey task will all likely provide better discriminant validity without compromising convergent validity.

Altogether, this speaks further to existing concerns about the generalizability of scales that have been validated exclusively with college students (Peterson, 2001). The same survey design elements likely to increase satisficing in lower ability samples either may not increase task difficulty for higher ability samples or simply may not increase their satisficing behaviors. For example, high ability respondents may be in the habit of looking past design problems. Taking this reasoning a step further, there might even be different sets of optimal characteristics for differently able groups, an idea that reframes the current practice of holding most research factors constant. Rather than administering the same test to everyone, slightly different forms of the same survey could produce more truly comparable results by removing known sources of satisficing error for lower ability samples. Many surveys offer different adult, adolescent, and/or child directed forms; it may be that there should be a wider range within the adult category, based on ability. Future research should investigate the effects of ability differentiated surveys.

New Framework. The current research demonstrates the damage of survey satisficing on results at several levels and it provides important insight into how researchers might improve survey designs, which possess characteristics known to trigger or

aggravate satisficing behaviors. Based on the interaction effects that this research found between correlates of survey satisficing—prior satisficing behavior, ability, motivation, and task difficulty—in predicting validity, we suggest that researchers be mindful of how their own choices affect survey satisficing.

As predictors of survey satisficing continue to be identified and their relationships better understood, researchers will have the opportunity to correct inflated reliabilities and validity measures by making survey design changes. These findings should not be perceived as a threat to well-established, reliable, and valid surveys as they simply provide a new framework from which to identify and reduce survey-satisficing error. The real promise of this research is a means of gleaning cleaner, less error-riddled research results from self-report survey data.

Limitations

One limitation for the present study is that we had a very rough operationalization of cognitive ability: education level. Of course, the university sample were all undergraduates, so we only compared the MTurk samples with and without college degrees to evaluate ability. However, as mentioned previously, even the MTurk sample without college degrees are computer savvy enough to participate in the study. Therefore, the results concerning ability should not be viewed as being definitive. Moreover, the differences between the student and MTurk samples were far greater, so we wish to emphasize the findings concerning motivation over ability.

A second limitation is that we examined only seven surveys in the first part and two surveys in the second. In particular, we examined discriminant validity using the set of interitem correlations between the items of two 15-item scales—one for extraversion and one for neuroticism—and we measured internal consistency reliability and convergent validity as the set of interitem correlations between items within each of these two scales. Additionally, we did not test for any associations with criterion-related validity.

Last, a major limitation is that we only measured survey satisficing with nondifferentiation. If nondifferentiation perfectly captured survey satisficing tendencies, then we would expect that label format and sample type would have significantly predicted nondifferentiation in main effects and interaction effects in the first part of the study; then, we would expect that only nondifferentiation (not label format or sample type) would predict inter-item correlation size in the second part of the study. Given the reality that nondifferentiation imperfectly measures satisficing, it makes sense that the determinants of survey satisficing (label format and sample type)—which presumably affected satisficing to a greater extent than was captured by nondifferentiation—would predict the psychometric outcomes of satisficing in both main effects and in interactions with nondifferentiation. However, due to the imperfect operationalization of satisficing as nondifferentiation, we may only assume that the above explanation adequately characterizes the causes and effects of survey satisficing in the present study.

Thus, the present findings must first be replicated with different measures of satisficing and different forms of validity before we can draw any definitive and generalizable conclusions. On the other hand, the numerous strengths of the present research—such as a relatively large sample, powerful statistical techniques, an experimental design, and surprisingly large effect sizes—make this study an important contribution to the literature on the predictors and effects of survey satisficing.

Conclusion

These research findings provide important insights into some fundamental aspects of survey satisficing, many of which can be influenced or controlled by researchers. First, we found evidence that for-pay survey sites, by design, provoke survey-satisficing behavior, apparently regardless of a subject's ability or task difficulty. Additionally, the conditions assumed to increase the propensity of satisficing behavior negatively affected psychometric outcomes; and they were especially impactful for those already prone to satisficing behavior. These results strongly support the hypotheses that our measures of previous satisficing behavior (nondifferentiation), less ability (participants from MTurk without degrees), lowered motivation (participants from MTurk in general), and task difficulty (endpoint labeled response scales) were all predictive of inflated interitem correlation sizes. The good news for researchers is that it is possible that small design improvements could lead to more valid and potentially more replicable survey results.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partly supported by Grant ECCS-1405173 from the National Science Foundation to the first author.

Notes

1. We similarly ran an ANOVA with sample characteristics, the response-scale label format, and the number of response categories predicting average within-subject standard deviations across the seven questionnaires (see Lelkes et al., 2012). These results were similar to those reported for the logistic regression.
2. As suggested by an anonymous reviewer, we performed these same analyses on the interitem partial correlations, controlling for gender, age, and ethnicity, but the results were very similar to those reported. We also tested for effects of response scale length, but the effect was weak. For simplicity, we do not report the results for either of these analyses.

References

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, *53*, 182-200. doi:10.1007/s11162-011-9251-2

Bass, B. M., Cascio, W. F., & O'Conner, E. J. (1974). Magnitude estimation of expressions of frequency and amount. *Journal of Applied Psychology*, *59*, 313-320. doi:10.1037/h0036653

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

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306-307. doi:10.1207/s15327752jpa 4803_13

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105. doi:10.1037/h0046016

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 21-56. doi:10.1016/j.chb.2013.05.009

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*, 980-1008. doi:10.1037/a0035661

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One*, *8*(3), e57410. doi:10.1371/journal.pone.0057410

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment*, *49*, 71-75. doi:10.1207/s15327752jpa4901_13

Dobson, K. S., & Mothersill, K. J. (1979). Equidistant category labels for construction of Likert-type scales. *Perceptual and Motor Skills*, *49*, 575-580. doi:10.2466/pms.1979 .49.2.575

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213-224. doi:10.1002/bdm.1753

Hamby, T. (2015). *An examination of the relationships between response scale length, label format, reliability and validity* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3709723)

Hamby, T., & Ickes, W. (2015). Do the readability and average item length of personality scales affect their reliability? Some meta-analytic answers. *Journal of Individual Differences*, *36*, 54-63. doi:10.1027/1614-0001/a000154

Hamby, T., & Levine, D. (2016). Response-scale formats and psychological distances between categories. *Applied Psychological Measurement*, *40*, 73-75. doi:10.1177/01466216 15597961

International Personality Item Pool. (2001). *A scientific collaboratory for the development of advanced measures of personality traits and other individual differences*. Retrieved from http://ipip.ori.org/

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory—Versions 4a and 5a*. Berkeley: University of California at Berkeley, Institute of Personality and Social Research.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236. doi:10.1002/acp.2350050305

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

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*, 526-538. doi:10.1086/269128

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Program Evaluation*, *70*, 29-44. doi:10.1002/ev.1033

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (pp. 263-314). San Diego, CA: Elsevier.

Lelkes, Y., Krosnick, J. A., Marx, D. M., Judd, C. M., & Park, B. (2012). Complete anonymity compromises the accuracy of self-reports. *Journal of Experimental Social Psychology*, *48*, 1291-1299. doi:10.1016/j.jesp.2012.07.002

Lennox, R. D., & Wolfe, R. N. (1984). Revision of the Self-Monitoring Scale. *Journal of Personality and Social Psychology*, *46*, 1349-1364. doi:10.1037//0022-3514.46.6.1349

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly*, *72*, 914-934. doi:10.1093/poq/nfn050

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*, 58-88. doi:10.1086/297739

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867-872. doi:10.1016/j.jesp.2009.03.009

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184-188. doi:10.1177/0963721414531598

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411-419. doi:10.2139/ssrn.2100631

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023-1031. doi:10.3758/s13428-013-0434-y

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*, 450-461. doi:10.1086/323732

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology*, *98*, 194-198. doi:10.1037/a0030767

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172-179. doi:10.1016/j.jtbi.2011.03.004

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304-307. doi:10.1016/j.chb.2014.11.004

Schwarz, N., & Strack, F. (1985). Cognitive and affective processes in judgments of subjective well-being: A preliminary model. In H. Brandstatter & E. Kirchler (Eds.), *Economic Psychology* (pp. 439-447). Linz, Austria: Tauner.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129-138. doi:10.1037/h0042769

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

Watson, D., O'Hara, M. W., Kotov, R., Simms, L. J., Chmielewski, M., McDade-Montez, E. A., . . . Stuart, S. (2007). Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychological Assessment*, *19*, 253-268. doi:10.1037/1040-3590.19.3.253

Zhang, C., & Conrad, F. G. (2013). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*, 127-135. doi: 10.18148/srm/2014.v8i2.5453#sthash.L6o1Pduf.dpuf