



Using the Theory of Satisficing to Evaluate the Quality of Survey Data

Citation

Barge, Scott, and Hunter Gehlbach. 2012. "Using the Theory of Satisficing to Evaluate the Quality of Survey Data." *Research in Higher Education* 53 (2) (March 30): 182-200. doi:10.1007/s11162-011-9251-2. <http://dx.doi.org/10.1007/s11162-011-9251-2>.

Published Version

doi:10.1007/s11162-011-9251-2

Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:11384947>

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)

Using the Theory of Satisficing
to Evaluate the Quality of Survey Data

Scott Barge
Goshen College
1700 S. Main Street
Goshen, IN 46526
scott_barge@mail.harvard.edu
Ph. 617-513-7106

Hunter Gehlbach
Harvard University
Graduate School of Education
Longfellow 328, 13 Appian Way
Cambridge, MA 02138

Introduction

Between administrative surveys, course-related questionnaires, end-of-term course evaluations, and student-initiated projects, (not to mention marketing from external groups) members of campus communities are regularly asked to complete a “short survey.” Students, faculty, and staff within higher education are inundated with survey requests. Thus, it seems inevitable that the resulting datasets contain at least a few responses from those who were less than optimally attentive and/or motivated in responding. Particularly in the context of web-based surveys, respondents may be distracted by a simultaneous Facebook conversation, listening to TV or music in the background, interrupted by a visitor, rushing to squeeze the survey in between two other commitments, or simply treating the survey as the means to a reward. By the same token, the resulting data set is also likely to include responses from thoughtful, motivated, deliberate participants. Separating out these sub-optimal respondents is pivotal to understanding fully one’s data.

As a framework for our investigation, we use the theory of satisficing—the idea that people often engage in sub-optimal decision-making strategies to conserve cognitive effort (Simon, 1957; Simon and Stedry, 1969)—and focus on its more recent application to survey research (e.g., Krosnick, 1991). Using this framework we demonstrate analytic options for assessing respondents’ conscientiousness in giving high fidelity survey answers. Specifically, we operationalize satisficing in the context of web-based surveys as a series of measurable behaviors and show how satisficing metrics can be computed for each survey respondent. These metrics can assess the sensitivity of survey results to variation in respondent effort. As we will

demonstrate, answers to substantive questions of interest in higher education can differ markedly depending upon respondents' level of satisficing.

Background & Context

Surveying on College and University Campuses

As surveys proliferate on college and university campuses, data are gathered from virtually all student populations, including prospective and current students, graduating seniors, and alumni as well as faculty and staff. Beyond institutionally sponsored surveys, research groups, course instructors, student research teams, and student organizations regularly distribute questionnaires to collect data. An informal review of several large research universities' campus online survey calendars reveals, on average, more than 10 official surveys in the current academic year; furthermore, there are undoubtedly surveys that are not included on the formal calendar.

Data from these surveys are used for a wide array of purposes. In some cases, survey administration is an essential part of compliance—perhaps for state or local subsidies, or to fulfill federal grant requirements. In other cases, survey data drive institutional research and planning (see, for example, Gonyea & Kuh, 2009). At some institutions, these data play an important role in accreditation—providing indirect measures of student learning, satisfaction, and other key outcomes (Snover, Terkla, Kim, Decarie & Brittingham, 2010). Given the current emphasis on accountability throughout education, surveys are only likely to grow more common and more important.

In addition, survey data are central to academic research on topics within higher education. In 2010, over half of the articles from *Research in Higher Education* (55%) relied on survey data in

some way. In some cases, these data are gathered using investigator-designed instruments specific to the study. Others make use of existing large-scale national datasets such as those available from Indiana University's National Survey of Student Engagement (NSSE) or UCLA's Higher Education Research Institute (HERI).

Survey data also play a significant role in the broader public discourse about higher education. Even a cursory review of mainstream publications such as *The Chronicle of Higher Education* reveals the frequency with which survey data capture media attention. For example, in 2010 alone, coverage of topics ranging from concern about college costs (Marchand, 2010) to online racial/ethnic bias (Laster, 2010) and student usage of Wikipedia (Miller, 2010) drew upon survey findings as the cornerstone of the story. In sum, survey data play a central role in higher education through decision-making, research, and the shaping of public opinion.

Taken together, these trends present a paradox for higher education. On one hand, data from surveys increasingly are used for evidence-based decision-making, academic scholarship, and informing public opinion. On the other hand, as the number of surveys increase, each one faces stiffer competition for respondents' time and attention. Thus, data quality may be declining as these data are playing increasingly important roles. This problem potentially adds an additional layer to the questions currently being asked about the validity of student surveys used in higher education (Porter, 2011).

Quality of Survey Data

Given this paradox, it stands to reason that, at least some respondents on some portion of survey items are not fully attentive. Researchers face the challenge of determining whether, and to what

extent, their survey data are contaminated with sub-optimal responses. While scholars occasionally report having checked their data for problems such as non-differentiation (i.e., “straight-line” responding), consistent, systematic reports of satisficing behaviors are essentially non-existent.

Survey response theory provides a framework for thinking more systematically about identifying low-quality responses. We can identify problematic respondents as those who *fail* to adequately engage in one or more of the stages of cognitive processing: (1) accurately interpret the meaning of the question being posed; (2) search memory for relevant information; (3) summarize information into a single judgment; and (4) convey that judgment in the most accurate way given the response options (Tourangeau, Rips, & Rasinsky, 2000).

Satisficing as a Theoretical Framework

The theory of *satisficing* offers a useful framework for exploring sub-optimal survey responses. The concept of satisficing was originally used to describe human behavior in cases when consumers’ decisions did not maximize personal gain (Simon, 1957; Simon and Stedry, 1969). These authors suggested that when required to process a demanding amount of information, some people only invest enough energy to make a *satisfactory* decision rather than one that *optimizes* benefits from the decision. The theory has been adapted to the field of survey research as a framework within which a variety of specific undesirable respondent behaviors might be understood (Tourangeau, 1984; Krosnick, 1991). In short, survey respondents satisfice when they fail to fully engage in one or more of the four stages of cognitive processing, thus degrading the integrity of their survey responses.

Respondents might satisfice through a variety of strategies. Krosnick (1991) discusses several forms: selecting the first response option that seems reasonable, agreeing with assertions (acquiescence bias), endorsing the *status quo*, not differentiating among response options for items using the same response scale (e.g., responding in a straight-line all the way down the page), selecting “don’t know” or “not applicable,” and randomly selecting a response. In addition, respondents might skip items or quit the survey early. For online surveys (with mechanisms to time response latencies), rushing through surveys is another observable form of satisficing.

Many surveys include information that can be used to determine the extent to which respondents engaged in satisficing behaviors. Non-differentiation can be identified easily in both online and paper-based datasets. The number of items skipped and the portion of the survey left unfinished are similarly straightforward to determine. Online surveying tools often afford a great deal of additional information about the respondents’ online session or sessions. Depending on the survey tool, survey start and end times, page times, and times between clicks can often be captured along with the response data. In nearly all instances of paper- and online-administration surveys, researchers have at their disposal at least a couple of means of identifying satisficing behaviors. Because the goal of this paper is to illustrate and evaluate the impact of different types of satisficing, we have chosen to work with data from two online surveys so that we might incorporate a broader range of techniques. We first describe how four metrics of satisficing can be computed. We then illustrate how the metrics can be aggregated into a satisficing index and used to evaluate the impact of satisficing on survey results.

Specifically, across these two studies we address the following research questions:

- RQ 1: To what extent did respondents engage in satisficing behaviors?
- RQ 2: Are key single-item results such as satisfaction sensitive to the effects of satisficing?
- RQ 3: Does satisficing adversely impact the reliability of scales?
- RQ 4: Does satisficing impact the association between variables or scales?

Research Design

Data

We use two large-scale survey datasets to develop satisficing metrics and investigate our research questions. Several differences between these two samples increase the likelihood that our results will generalize to a range of other surveys. While one survey was from Eastern Europe and was administered both in English and Lithuanian, the second survey was from the United States and administered in English. The Eastern European survey was administered at 15 colleges and universities, while the U.S. survey data are from a single institution. The former survey was of only moderate length, while the latter survey, as a summative evaluation of undergraduate educational experiences, was quite lengthy. One final difference between the surveys is that respondents to the Eastern European survey did not receive an incentive for completing the instrument, while U.S. students who finished the survey were able to select from two compensation options.

The first dataset—Survey 1—was gathered as part of a multi-university study of student academic engagement in Lithuania. The survey contained roughly 75 items (depending upon the

survey form) addressing various aspects of students' academic experiences. On average, the survey took students around 8-10 minutes to complete using an online surveying tool. Study participants received neither incentive nor compensation for their participation. The final sample included 20,952 responses, which represents a 43% response rate.

The survey covered topics such as students' interaction with peers and faculty members, their class participation, academic workload, time allocation, and satisfaction. The academic engagement items listed 23 specific activities, such as "Contributed to class discussions" and asked respondents to indicate how often they did each activity on a four-point response scale. These items together formed an engagement scale ($\alpha = .85$). Another 17-item scale ($\alpha = .91$) asked students to rate, using a five-point response scale, the extent to which their experiences at their universities contributed to their development of specific skills and competencies such as critical thinking and working effectively with others. Additional survey items incorporated a 5-item higher-order thinking scale ($\alpha = .70$), which used a four-point response scale, as well as questions about how many hours per week they spent on specific tasks, and their overall satisfaction.

The Survey 2 data come from a highly selective U.S. university's regular survey of their undergraduate seniors. This survey contained more than 250 items, and students averaged 25 minutes to complete the questionnaire. Students completing this survey were offered either a \$15 cash incentive or a \$20 donation in their name to the senior class gift. The final sample, representing a response rate of 65%, was 755.

This survey served as a summative evaluation of respondents' undergraduate experiences. Using a four-point response scale, it asked students about their satisfaction with academic quality, campus services and key aspects of their major program of study. This latter set of questions together formed a 15-item college major satisfaction scale ($\alpha = .80$). Students further completed an extensive 25-item bank listing specific skills and competencies. This bank asked students to indicate how important each skill was to them using a four-point response scale, and the extent to which their experiences at the university contributed to their development of the skill, also using a four-point response scale. The development portion of this bank included five scales for: critical thinking and communication (9 items; $\alpha = .81$), arts and humanities (6 items; $\alpha = .75$), scientific/quantitative reasoning (3 items; $\alpha = .66$), independent/confident self (4 items; $\alpha = .68$), and interpersonal skills (3 items; $\alpha = .65$). The survey also included items about students' future plans, financing of their undergraduate studies, and participation in various activities.

Procedures & Measures

For each survey we computed four specific satisficing metrics, each based on observable sub-optimal respondent behaviors: exiting the survey early, non-differentiation, skipping items, and rushing. These four measures are derived from features of each data set. The first three measures apply to all surveys—paper and online, while the fourth metric is more commonly available with online surveys. Here we detail each metric we developed for this study, describing the specific behavior behind the metric and our approach to computing it.

Early Termination

The first metric we use, “early termination,” is the percentage of items that the respondents do not complete because they exit the survey without finishing and do not return. It is more common in longer surveys, and often occurs after a page break. We calculate this metric as a percentage:

$$\text{Early Termination}_i = \frac{\text{Number of items unfinished}}{\text{Total number of items}} \quad (1)$$

Non-Differentiation

Non-differentiation, our second metric, occurs when survey respondents select the same response option across a set of similarly structured and/or related items. Consider as an example the 25-item survey bank from Survey 2, asking about skill and competency development, using a four-point response scale (Weaker now, No change, Stronger now, Much stronger now). A respondent who selected ‘Weaker now’ for all 25 items could be considered to be non-differentiating. In this study we do not consider banks of less than five items, and offer five as a rule-of-thumb minimum. However, as we discuss later, determining which banks should be incorporated into non-differentiation analysis depends on a number of factors. To identify an instance of non-differentiation j , we computed the respondent-level standard deviation for the items in each bank in which non-differentiation was possible.

$$\text{Standard Deviation}_j = \sqrt{\frac{S(x - \bar{x}_j)^2}{n - 1}} \quad (2)$$

Where x is the specific coded value for the item (here ‘Weaker now’ is coded as 1), \bar{x}_j is the respondent-level mean of the response values, and n is the number of items. In the given example with ‘Weaker now coded as 1, the respondent-level standard deviation across the 25 development items would be 0 since there was no variability in response selection across the bank. As such, we considered any respondent with a standard deviation of 0 across the 25 items of the time bank to be non-differentiating. This approach is particularly stringent; we later discuss more nuanced approaches. Because satisficing has not been looked at systematically, we wanted to use a more restrictive definition to make sure it was important even when we focused only on its extreme forms. Thus, to some degree our findings can be looked at as the lower bound of how problematic satisficing behaviors can be.

[FIGURE 1 ABOUT HERE]

The surveys in this study contain multiple item banks on which respondents could engage in non-differentiation. To compute the non-differentiation metric, we aggregated the instances of non-differentiation as a percentage of the opportunities for non-differentiation.

$$\text{Non-differentiation}_i = \frac{\text{Instances of non-differentiation}}{\text{Opportunities for non-differentiation}} \quad (3)$$

Thus, if the survey contained three banks on which it was possible to non-differentiate, and respondents did so on two banks, their value for the metric would be $2/3 = .67$. In this calculation, we took into account only those banks that the respondent *viewed*. For respondents who exited the online survey before viewing the final bank and who non-differentiated on both of the banks that they responded to, their value for the metric would be $2/2 = 1.0$.

Skipping Items

Next, we create a metric for skipping items. We define this metric specifically as respondents failing to answer a question for which an answer is possible and applicable. Thus, we set aside any item for which response was conditional on a prior item (e.g., “If you replied yes on the last item, ...”) and any items that were relevant only to a subset of respondents. Further, we incorporate only items from the portion of the survey the respondent completed. We calculate the metric as a percentage:

$$\text{Skipped}_i = \frac{\text{Number of applicable items left blank}}{\text{Number of applicable items seen}} \quad (4)$$

Rushing

The final satisficing metric we use is derived from the speed with which the respondent works—or rushes—through the survey. We define rushing as speeding through the survey at a sufficiently fast rate that thoughtful answers to each item is highly unlikely. To calculate the metric we first compute a seconds-per-item rate (SPI) for each respondent:

$$\text{SPI}_i = \frac{\text{Number of seconds on survey}}{\text{Number of items completed}} \quad (5)$$

From the full sample of respondents, we then determine the modal value of SPI. We opt here for the mode rather than mean or median because these latter measures are influenced by the presence of extreme or nonsensical values (e.g., significant numbers of respondents who took a break from the survey and returned to it days later). In these cases, the SPI estimates are invalid

and their influence on the median and mean are problematic. By contrast, the mode captures the most common rate at which respondents proceeded through the questionnaire.

For respondents with a per-item rate *lower* than the modal value, we then compute the rushing metric in reference to the *fastest* rate (minimum SPI):

$$\text{If } SPI < \text{Mode}(SPI), \text{ Rushing}_i = \frac{\text{Minimum}(SPI)}{SPI_i} \quad (6)$$

$$\text{If } SPI \geq \text{Mode}(SPI), \text{ Rushing}_i = 0$$

Thus, the respondent in the sample who completed the survey in the least amount of time has a rushing value of 1, and those who were slightly faster than the modal speed have a rushing value just greater than 0. By computing this metric as a rate rather than a measure of time-on-survey, we account for the fewer number of items seen by those who terminate early or skip items.

Data-Analytic Plan

To address the research questions of this study we use the metrics independently and as a combined index. Since each metric has values between 0 and 1, we construct a satisficing index by adding the four metrics together. At the conceptual level, it makes sense to combine these four behaviors as observable forms of satisficing. Unlike a traditional psychometric scale, in which indicators of the same construct should correlate positively, the behaviors in this index will not necessarily correlate positively, as we discuss in our findings.

To illustrate the impact of satisficing behaviors more clearly, we also collapse the continuous metric to create five different groups of “satisficers” (or “non-satisficers”). The first group

consists of respondents with values of 0 on the index—those who do not satisfice. The remaining four groups are defined by quartiles of the index: weak satisficers are those with index values in the first quartile, the second quartile forms the moderate satisficers group, and the third and fourth quartiles are termed strong, and very strong satisficers.¹

For research question one, we report the prevalence of each type of satisficing behavior for both study samples. For the second research question, we investigate the distributions of several survey items that were similar on both questionnaires, focusing on differences between non-satisficers and very strong satisficers. To answer research question three, we compare the internal consistency reliability (using coefficient alpha) of several scales across different groups of satisficers. Finally, for the fourth research question we compare the correlations of the scales from research question three across different groups of satisficers.

Findings

To what extent did respondents satisfice?

In both Survey 1 and Survey 2, the majority of survey respondents engaged in at least one of the satisficing behaviors we focus on. Among the Lithuanian university students of Survey 1, 61% engaged in at least one form of satisficing, 13% in at least two forms, 1% in three forms, and less than 1% in all four forms. Skipping items was most common, with 39% opting not to answer at least one item.

¹ Due to the dangers of artificially dividing continuous data (Maxwell & Delaney, 1993), many of our analyses focus on comparing non-satisficers and very strong satisficers. We do, however, present results across all four quartiles.

Students at the U.S. private research university were more likely to engage in one of these forms of satisficing. In Survey 2, 81% of respondents engaged in at least one satisficing behavior, 41% in at least two forms, 10% in at least three forms, and 1% in all four forms. Again, skipping was the most common, with 69% of students skipping at least one item. In both studies, rushing was the second most common of the satisficing behaviors. In Survey 1, early termination and non-differentiation were the third and fourth most common behaviors, respectively. Among Survey 2 respondents, this ordering was reversed—students were more likely to non-differentiate than exit the survey without completing it. See Table 1 for further detail on the prevalence of satisficing behaviors, descriptive statistics of the satisficing metrics, and the correlations between metrics.

[TABLE 1 ABOUT HERE]

Though each individual metric ranges from 0 to 1, by definition the index cannot have values of 3 or 4. That is, it is impossible for a respondent to skip 100% of the survey (for a value of 1 on the skip metric), terminate early so as to not see 100% of the items, and then non-differentiate on 100% of the items. In Survey 1, the mean value for the satisficing metric was 0.175 (std. dev. = .289), with a maximum value of 2.57. In Survey 2, the mean value was .217 (std. dev. = .279), with a maximum value of 1.65. Table 2 presents means of each metric, and the index itself, by satisficing group (i.e., by quartile of the satisficing index). On both surveys, the metrics and the overall index are positively skewed, suggesting that the majority of satisficers engage in the specified behaviors to a limited extent, while a minority are more severe satisficers.

[TABLE 2 ABOUT HERE]

Are key single-item results such as satisfaction sensitive to the effects of satisficing?

Survey 1 included several questions that are common to undergraduate surveys: an item on overall satisfaction, a time diary question on the number of hours per week spent preparing for class, and a self-report of student grade-point average. The distribution of respondents across response categories for these items differed across the quartiles of the satisficing index. In Table 3 we present these distributions. The satisfaction item asked respondents to evaluate their entire educational experience at their particular institution. The five response categories ranged from “Very bad” to “Very good.” Satisficers differed from non-satisficers in how satisfied they were; note in particular the difference at the lowest level of satisfaction (Very bad). As Table 3 indicates, for non-satisficing students this percentage is less than 1%. Among very strong satisficers, four times as many students (4%) evaluated their experience as “Very bad.” Alone, this finding might reflect real differences between sub-populations of students, as it is reasonable to expect that students having poor experiences might be less likely to pay careful attention to a survey. However, when this difference is considered alongside differences on other items, inconsistencies emerge.

[TABLE 3 ABOUT HERE]

The differences between non-satisficers and very strong satisficers on the preparation time and grade-point average items are even more pronounced. Once again, note the differences at the extremes of the response options. While 1.5% of non-satisficing students indicate they spend zero hours per week studying, the figure is 6.5% among very strong satisficers. With regard to grade-point average, 2.4% of non-satisficers report an average of 10 on the European grading scale (equivalent to A on the common U.S. scale). Among very strong satisficers, the percentage of students who report an average grade of 10 jumps to 76% -- it is hard to imagine that this

nearly 75 percentage point increase reflects the real differences between these populations of students. Specifically, it seems hard to reconcile that very strong satisficers are more likely to be dis-satisfied and report less time studying while being over 30 times more likely to report *perfect* grade-point averages.

Survey 2 also included questions about overall satisfaction, time spent studying and self-reported grade-point average. Though the differences in distribution for satisfaction do not track with those of Survey 1, respondents in Survey 2 exhibit similar differences between non-satisficers and very strong satisficers on the other items (See Table 2). Replicating the perplexing finding described above, very strong satisficers are more likely to report zero hours studying than their non-satisficing peers (6.4% vs. 1.3%); yet, they are *more* likely to report average grades of A (19%) than are non-satisficing respondents (9%).

In both surveys, the more pronounced differences in distributions across response anchors for the satisfaction and study time items exist among very strong satisficers. This mirrors the positive skew of the metrics themselves. That is, non-satisficers, weak satisficers, moderate satisficers, and strong satisficers are relatively similar in their distribution across the response options for these two topics. Among very strong satisficers, which have more extreme values on the metrics and index, the distributions shift. Interestingly, though, this pattern does not hold for grade reporting. All satisficers were more likely to report “A” as their average grade.

Does satisficing adversely impact the reliability of scales?

In both studies, the survey instruments included banks of items designed to form scales. In Survey 1, the three scales captured respondents’ self-reported academic engagement, use of

higher-order thinking skills in coursework, and development on various outcomes. The Survey 2 scales also measured self-reported development along a variety of dimensions as well as respondents' satisfaction with their experience in their particular major area of study.

In Table 4, we present the internal consistency reliability (using coefficient alpha) for each scale, across all five levels of satisficing. Across both samples, scale reliabilities tend to decline among weak satisficers, and then increase as the degree of satisficing increases to very strong. In eight of nine cases, the differences in reliability between non-satisficers and very strong satisficers are statistically significant ($p < .01$).² The increase in reliability moving from non-satisficers to very strong satisficers tends to be steadier for longer scales (i.e., scales consisting of more items).

[TABLE 4 ABOUT HERE]

Naturally, respondents who non-differentiate as they respond to the scale items drive these increases in reliability in large part. To test the extent to which this behavior alone accounts for the increase in reliability, we set aside those respondents who non-differentiated on each respective scale and re-calculated the reliability coefficients for the "very strong" satisficers. The reliability increases are less pronounced, but remain. For example, for the engagement scale in Survey 1, the reliability coefficient increases from .841 to .857 ($W=1.11, p=.03$) rather than to .896 ($W=1.53, p<.01$) when non-differentiating respondents are included. Reliability of the

² Here we use Feldt's (1969) W statistic to test the significance of differences in coefficient alpha. In evaluating the impact of satisficing on scale reliabilities (RQ 3) and scale correlations (RQ 4) we conduct statistical tests to compare the differences between non-satisficers and very strong satisficers. Our decision here is motivated by the pronounced nature of the differences between these two groups, the positive skew of the satisficing metrics and index, and a desire to avoid the accumulation of Type I error that would result from conducting tests for each difference.

Survey 1 development scale increases from .901 to .921 ($W=1.25, p=.01$), as opposed to .974 ($W=3.81, p<.01$).

Does satisficing influence the association between scales?

To explore the impact of satisficing on the statistical relationships between constructs, we calculated the correlations between several scales in both surveys. In most cases, these are scales that should be positively related to one another. For example, academic engagement should be positively associated with use of higher-level thinking skills in the classroom. Likewise, we might logically expect self-reported critical thinking abilities to be positively associated with development of scientific and quantitative reasoning skills. In one case, we also test the relationship between two scales that are not expected to correlate (satisfaction with major and development of interpersonal skills).

In Table 5 we present selected scale correlations across the five levels of satisficing. In nearly all cases, the scale correlations are higher among very strong satisficers than for their non-satisficing peers. These differences are statistically significant in five of six instances. The single exception to the general trend is the correlation between engagement and higher-level thinking in Survey 1, which is lower among very strong satisficers. The general trend of increases in correlation, however, is not monotonic. In four of the examples, the correlations decline before the more pronounced increases among very strong satisficers.

Here again, because of the strong impact of non-differentiation on scales, we set aside those respondents who non-differentiated on each respective scale and examined the correlations. In these cases, the increases generally did not persist. In four of the five instances, the scale

correlations were lower among very strong satisficers when the impact of non-differentiators was excluded.³ The single exception to this was the increase in correlation of the major satisfaction and interpersonal skills scales of Survey 2, which persisted even when non-differentiators were set aside.

[TABLE 5 ABOUT HERE]

Discussion

In this paper we have presented a theoretical framework and methodology for evaluating the quality of survey data. Using two survey data sets, we demonstrated these methods and illustrated the potential impact of satisficers on data quality. Several subtleties in the results deserve further discussion. We begin by interpreting our findings and describing the subsequent implications. We then move to important methodological considerations, and finally offer conclusions.

Interpretation of Findings

Prevalence of Satisficing

Our analysis from both Survey 1 and Survey 2 confirm that respondents did indeed engage in all four of the satisficing behaviors we investigated. In fact, on both surveys, a majority of respondents engaged in at least one satisficing behavior. Clearly, not all respondents are dedicating the time and attention to the survey that researchers would like. Though not supportive of causal inferences, our data suggest several hypotheses about how and why

³ Here we did not consider the higher-level thinking and satisfaction scales of Survey 1, as we did not feel they were long enough to justify computation of a non-differentiation metric.

respondents satisfice. Most respondents selected one, or perhaps two ways to reduce the effort they dedicate to the survey. Rarely did respondents engage in all four forms of satisficing. As we allude to above, traditionally, common indicators of a construct would be expected to correlate with one another. Yet the focal behaviors in our satisficing index represent distinct ways in which survey respondents might inadequately engage in the necessary four stages of cognitive processing. Some are positively associated with one another—such as non-differentiation and rushing, which correlate positively $r=0.1$ ($p<.05$) in both studies. Other behaviors, however, are likely to be *negatively* associated. For example, respondents who lose patience as a survey grows long might start to skip items or exit the survey, but they cannot do both. We observed this negative association in both studies.

As we noted in presenting descriptive statistics for the satisficing metrics, the majority of respondents engage in satisficing behaviors to a limited extent. The resulting positive skewness highlights an important area for further research and attention. Specifically, though most respondents engage in at least one satisficing behavior, what is the threshold at which satisficing has negative and statistically significant consequences on the data? Are researchers most likely to find that very strong satisficers are primarily responsible for observed problems? A deeper look into these questions will be beneficial to the field generally, but we would also note that the distributions of satisficing metrics are likely impacted by context and, as such, eschew the application of universal guidelines.

As a final observation on the prevalence of satisficing, we noted in additional analyses that satisficing increased as respondents progressed through the survey. For example, on Survey 1, roughly 1% of respondents non-differentiated on the first bank of questions, while on the last

bank of questions the proportion more than doubled to 2.5%. In Survey 2 this trend was similar. At the first opportunity to non-differentiate, 2% of respondents did so, while on the last opportunity 5% engaged in this behavior. Survey fatigue is a logical cause for this, though an experimental design would be needed to evaluate this explanation more definitively.

Impact on Single Items

Our findings suggest that satisficing is associated with differences in the distributions of respondents on key single-item indicators. This appears to be true with multiple types of items—attitudinal (satisfaction), behavioral (time spent studying), and factual (self-reported GPA). In some cases—hours spent studying and GPA—the problems that emerge among satisficing respondents are relatively clear-cut. Satisficing respondents are providing highly unlikely responses that degrade the overall quality of the data. In other cases, the nature of the relationship between satisficing behaviors and the impact on distributions is more difficult to disentangle. For example, are satisficing respondents more likely to indicate their dis-satisfaction because of idiosyncrasies of the question layout, or are respondents who are truly dis-satisfied with their experiences more likely to satisfice? In the case of the satisfaction items we worked with, one possible explanation for the satisfaction inconsistencies is the placement of the satisfaction question in each survey. On Survey 1, the satisfaction question was at the end of the survey, followed only by demographic items. On Survey 2, the satisfaction question was the first item on the survey. Given that satisficing tended to be more prevalent later in these surveys, we might assume that satisficing respondents had little or no impact on the responses to the satisfaction question on Survey 2. Regardless of the particular cause, the satisficing metrics facilitate investigation of these differences.

Ultimately, awareness of the possible impact of satisficing on single item results allows for more accurate and valid interpretation of findings. This becomes particularly important when institutions use key single items for evaluative purposes—perhaps gauging the impact of a particular student service, capturing the frequency of student collaboration, or measuring student satisfaction with faculty advising. In these situations, undetected satisficing reduces the validity of the measure and leaves room for erroneous conclusions. For example, reporting on the GPA items would suggest that 18% of students taking Survey 1 have perfect grades (rather than the 2.4% when satisficers are excluded) or that 20% of students taking Survey 2 receive all A's, when, in fact, among non-satisficers it is closer to 10%. Likewise, using the GPA items for disaggregation of other questions might lead to erroneous conclusions about the attitudes or opinions of “so-called” top students. One can imagine these findings leading to particularly troubling conclusions if researchers then proceeded to try to investigate the habits of these highly achieving students – perhaps finding that a disproportionately large number of them studied for zero hours per week.

Impact on Scales

Satisficing also appears to cause problems for scale reliabilities and the relationships between scales. In general, as satisficing becomes more pronounced, the reliabilities of scales increase, as do the correlations between scales. Much of this is a result of the impact of non-differentiation on scale reliabilities. As greater numbers of respondents opt for the same response anchor in a scale, the reliability coefficient approaches 1.0⁴. Even when we set aside those who non-

⁴ This is true so long as *all* respondents do not non-differentiate using the same response anchor (e.g., all respondents in the sample opting for 4 on a 4-point scale). If this were the case, the scale

differentiated and recomputed the satisficing index, we still find that reliabilities remain slightly inflated for satisficers. One possible reason for this persistent inflation could be respondents who, though they do not select the exact same response for the bank of items, insufficiently adjust their responses from one item to the next. In these cases, respondents might respond to the first item in the bank and then use their first response as an anchor from which they only minimally adjust as they complete the subsequent item. This “anchoring and adjusting” behavior might continue for the entire bank of similar items. The net effect would be spurious increases in scale reliability (though more subtle increases than in the case of non-differentiation).

Our results also point to satisficing’s potentially problematic impact on correlations between scales. In most cases, we found scales to be more highly correlated among very strong satisficers than among non-satisficers. This, like the impact of satisficing on scale reliability, is likely due in large part to the impact of non-differentiation. Particularly when respondents non-differentiate at one of the extreme ends of the response scale, the inter-scale correlations will be driven up. At its worst, this pattern can establish correlations that in fact do not exist among non-differentiating respondents. For example, in Survey 2, the correlation between students’ satisfaction with their major and their self-reported development of interpersonal skills among non-satisficers is not statistically significant at ($r = -.06$). As satisficing increases, however, this correlation becomes positive and moderately strong such that among very strong satisficers, it is statistically significant ($r = .42$).

variance would be zero, and Cronbach’s alpha cannot be computed. However, if *all* respondents non-differentiate and even *one* respondent opts for an alternate response anchor (e.g., 3 on a 4-point scale), alpha will be 1.0.

Given the potential for these types of spurious fluctuations in both reliability and inter-scale correlations, clearly the negative impact of satisficing on data quality extends beyond single items. As institutions use survey data to understand the associations between curriculum or programming and educational outcomes, correlations between psychometric scales are often foundational to analysis. As we have shown, satisficing can falsely inflate both the reliability and correlations in these types of scales. In the most extreme cases, satisficing respondents can negatively influence the data enough to introduce correlations where, in fact, none exist.

The Impact of Incentives

The ways in which respondents satisficed on Survey 2 raise questions about the use of incentives that seem important to explore in future research. Students who completed this survey—that is, who worked to the end of the instrument and clicked “submit”—were compensated with a \$15 credit to their student debit cards (or the option of a \$20 contribution to the class gift in their name). Skipping, rushing, and non-differentiation were decidedly more common in Survey 2 than Survey 1. Yet, early termination (a strategy that would require foregoing rewards) was more common on Survey 1. In a desire to earn the promised incentives, it is possible that Survey 2 respondents left items blank and/or simply clicked their way through the instrument, paying little attention to the substance of the questions.

For those conducting survey research on higher education populations, it seems crucial to know whether incentives sometime degrade individual survey responses even as they potentially boost overall response rates. Researchers often fixate on response rate as the key metric for determining whether or not a survey sample is sufficient for answering the research questions at

hand. Accordingly, incentives function as a relatively reliable means of driving up response rates (Dillman, Smyth, & Christian, 2009). What remains unclear, however, is whether unintended and perhaps negative consequences may result from using incentives. If it turns out that incentives can degrade item-level data quality under certain situations, many institutions may need to rethink their data collection plans.

Because our two survey populations differ along many dimensions, the differences we observe could result from a wide array of other factors as well. Nevertheless, the trends point to a crucial issue to examine in future randomized experiments. Such research might also include a third motivational context—requirements to complete surveys. As colleges and universities have become more dependent on survey data, many have required completion of certain surveys (e.g., course evaluations or graduation exit surveys). Though our study here does not incorporate data from a compulsory response context, we might hypothesize that satisficing behaviors similar to those present in an incentive context could plague compulsory surveys as well. The type of satisficing index we use here provides an important measure for facilitating this type of study.

Methodological Approach

Because our goals were to illustrate and measure satisficing, we calculated our metrics in straightforward ways. However, each metric offers the opportunity for added complexity and the potential for more nuanced measures as a result. Depending upon the specifics of the data set, each of our four metrics could be re-calculated in a more sophisticated form.

The bank-specific definition of non-differentiation need not be defined solely in terms of a respondent-level standard deviation of zero. Our choice to use zero reflects a stringent (possibly

overly stringent) definition of non-differentiation. In some contexts, it may be preferable to relax the definition such that a respondent is considered to non-differentiate if he/she selects the same response choice in all but one or two cases in the bank. For lengthier banks this approach might make more sense. Alternately, there might be banks of similar items in which it is conceivable that a fully attentive respondent would justifiably select the same response choice in all cases. Ultimately, decisions about which banks should be included in calculation of a non-differentiation metric should be informed by the content of the items, the specific context of the survey, and above all, the sound judgment of the researcher.

Our assumption in calculating both the percentage of items skipped and the percentage of items unfinished (i.e., early termination) is that we knew the point at which respondents exited the survey, if they did so. In both of our data sets we were able to ascertain this point using the last-answered item as a marker. If respondents' point of exit is ambiguous for several participants (e.g., several pages are blank or multiple sections have been skipped), a researcher interested in conducting a satisficing analysis could operationalize these two behaviors in a *single* metric that captures unanswered survey items. In such cases, the unanswered items metric would be calculated as the number of items unanswered divided by the number of items presented to the respondent.

In our analysis here, we opted for an omnibus metric of respondent speed that relied solely on start and stop time. In fact, many online surveying tools are capable of gathering page-specific start and stop times or even gathering time between clicks. With more detailed data come greater opportunities for more granular approaches to measuring rushing. For example, if a specific bank

of questions or scale is particularly important in addressing a research question the rushing metrics might focus solely on that sub-set of items.

Though not directly part of our investigation here, the impact of survey design on satisficing is an important consideration. Existing research offers many best practices for the construction and administration of surveys (e.g., Dillman, Smyth, & Christian, 2009). It is reasonable to expect that careful attention to instrument design might decrease the likelihood of satisficing. Although the two surveys we have incorporated in this study did not necessarily adhere to the principles of good survey design, we suspected that they represented typical surveys in higher education in many ways. However, another important area for future research – that is made possible by these types of satisficing metrics – is the investigation of ways that improving the design of these kinds of surveys might lead to a reduction in satisficing.

Finally, it is important to consider that the four forms of satisficing we incorporate into this study are not the only behaviors that would fit the theoretical definition. For example, researchers might tabulate the number of “don’t know” or “not applicable” responses as a satisficing metric. Other forms of satisficing, such as selecting the first response option that seems reasonable, or agreeing with assertions (acquiescence bias) would require a split ballot experimental design to measure.

Conclusion

As we have highlighted with this research, satisficing behaviors have a measurable impact on survey results. In the vast majority of cases, researchers have at their disposal a variety of metrics they can use to identify respondents who are engaging in satisficing behaviors. The calculation

of these metrics can be straightforward, and their utility in identifying problematic responses is apparent. Less clear is what researchers *should* do once satisficing respondents have been found. Offering any hard and fast rules for how to handle satisficers would invite myriad valid exceptions. At a minimum, however, we suggest researchers conduct key analyses with and without satisficers (or perhaps some subgroup of particularly strong satisficers) present in the data set. This will highlight any impact these problematic responses have on the data. If results with satisficers set aside are not satisfactory, it might be worthwhile to consider weighting responses inversely proportional to the degree of satisficing. As we have noted in describing the construction of the metrics, context and the sound judgment of the researcher are of the utmost importance in these types of analyses.

We present this article as a beginning, not an ending. Our hope is that by offering a theoretical framework and systematic methodological approach to evaluating the quality of responses, we advance the broader conversation about the quality of survey data. These data play an increasingly important role in compliance, decision-making, and accreditation for universities. They further continue to be vital to education research and remain influential in public discourse. As online tools make it easier for researchers to send out surveys, these surveys also draw less attention from respondents. These same online tools, though, afford researchers additional opportunities to evaluate respondent engagement. By identifying and isolating respondents who engage in sub-optimal survey response behaviors, researchers can evaluate the impact of these behaviors on their data.

References

- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method (3rd ed.)*. Hoboken, NJ: John Wiley, & Sons, Inc.
- Gonyea, R.M. & Kuh, G.D. (Eds.). (2009). Using NSSE in institutional research [Special issue]. *New Directions for Institutional Research*, 141
- Krosnick, J. A. (1991). Cognitive demands of attitude measures. *Applied Cognitive Psychology*, 5, 213-236.
- Kuh, G. D. (2001). The National Survey of Student Engagement: Conceptual framework and overview of psychometric properties. Bloomington, IN: Indiana University, Center for Postsecondary Research
- Laster, J. (2010, April 30). Black students experience more online bias than do whites. *The Chronicle of Higher Education*. Retrieved March 15, 2011, from <http://chronicle.com/blogs/wiredcampus/black-students-experience-more-online-bias-than-do-whites/23616>
- Marchand, A. (2010, January 21). Cost of college is a big worry of freshmen in national survey. *The Chronicle of Higher Education*. Retrieved March 15, 2011, from <http://chronicle.com/article/Cost-of-College-Is-a-Big-Worry/63671/>
- Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, 113(1), 181-190.

Miller, M. H. (2010, March 16). Students use Wikipedia early and often. *The Chronicle of Higher Education*. Retrieved March 15, 2011, from <http://chronicle.com/blogs/wiredcampus/students-use-wikipedia-earlyoften/21850>

Porter, S. R. (2011). Do College Student Surveys Have Any Validity? *The Review of Higher Education*, 35(1), 45-76.

Simon, H. A. (1957). *Models of man*. New York: Wiley.

Simon, H. A. and Stedry, A.C. (1968). Psychology and economics. In G. Lindzey and E. Aronson (Eds), *Handbook of Social Psychology*, 2nd edn, Vol. 5 (pp. 269-314). Reading, MA: Addison-Wesley.

Snover, L., Terkla, D. G., Kim, H., Decarie, L., Brittingham, B. (2010, June) NEASC Commission on Higher Education Requirements for Documentation of Assessment of Student Learning and Student Success in the Accreditation Process: Four Case Studies. In J. Carpenter-Hubin (Chair), *Charting Our Future in Higher Education*. 50th Annual Forum of the Association for Institutional Research (AIR), Chicago, IL.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. Jabine, M. Straf, J. Tanur, and R. Tourangeau (eds), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington, DC: National Academy Press.

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

Table 1. Descriptive statistics and correlations for satisficing metrics, by study.

Metric	Pct of Respondents	Mean (SD) of Metric ¹	Correlation matrix ¹		
			(1)	(2)	(3)
<i>Survey 1 (N=20,952)</i>					
Any satisficing behavior	61%				
Non-differentiation (1)	4%	.03 (.12)			
Skipping (2)	39%	.06 (.12)	.13***		
Early termination (3)	14%	.13 (.28)	-.02*	-.04***	
Rushing (4)	21%	.08 (.11)	.10***	-.21***	-.17***
<i>Survey 2 (N=809)</i>					
Any satisficing behavior	81%				
Non-differentiation (1)	19%	.21 (.17)			
Skipping (2)	69%	.06 (.13)	-.06		
Early termination (3)	9%	.05 (.18)	-.05	-.47**	
Rushing (4)	36%	.11 (.13)	.10*	-.17**	.04

¹Means and correlations of metrics calculated for those who engaged in any satisficing behavior (Survey 1 n=12,891; Survey 2 n=655).

* $p < .05$; ** $p < .01$, *** $p < .001$

Table 2. Means of satisficing metrics and satisficing index, by satisficing group.

Metric	Weak	Moderate	Strong	Very Strong
<i>Survey 1</i>				
Non-differentiation	.000	.002	.012	.094
Early termination	.001	.010	.009	.492
Skipping	.024	.055	.040	.113
Rushing	.000	.075	.187	.066
Satisficing Index	.024	.143	.248	.765
<i>Survey 2</i>				
Non-differentiation	.000	.110	.160	.280
Early termination	.000	.010	.000	.210
Skipping	.024	.041	.038	.134
Rushing	.000	.040	.190	.190
Satisficing Index	.024	.124	.256	.679

Table 3. Impact of satisficing on selected single items, scale reliabilities, and scale correlations, by study.

Item	Survey 1					Survey 2				
	Non-Satisficers	Weak Satisficers	Moderate Satisficers	Strong Satisficers	Very Strong Satisficers	Non-Satisficers	Weak Satisficers	Moderate Satisficers	Strong Satisficers	Very Strong Satisficers
Satisfaction										
Very bad	0.6%	0.8%	1.0%	1.6%	4.1%	2.1%	0.0%	2.9%	2.6%	0.0%
Bad	2.8%	3.0%	3.1%	4.5%	4.1%	3.2%	4.4%	6.6%	5.2%	2.7%
Fair	27.8%	31.5%	31.9%	29.4%	33.7%	10.6%	7.7%	10.9%	6.5%	5.5%
Good	54.3%	52.5%	50.7%	51.2%	45.6%	53.2%	55.8%	43.8%	44.8%	60.3%
Very good	14.5%	12.3%	13.4%	13.4%	12.7%	30.9%	32.0%	35.8%	40.9%	31.5%
	($\chi^2 = 150.82, p < .01$)					($\chi^2 = 24.237, p = .084$)				
Hours preparing for class										
0	1.5%	1.6%	2.4%	2.4%	6.5%	1.3%	1.6%	1.6%	1.3%	6.4%
1-5	26.1%	25.8%	28.3%	35.7%	34.0%	10.4%	3.7%	4.7%	8.4%	11.5%
6-10	25.6%	23.5%	25.7%	23.0%	24.3%	14.3%	14.4%	18.8%	16.2%	25.6%
11-20	29.9%	29.4%	29.2%	24.0%	21.5%	40.3%	43.3%	37.5%	45.4%	33.3%
More than 20	17.0%	19.7%	14.4%	14.8%	13.7%	33.8%	36.9%	37.5%	28.5%	23.2%
	($\chi^2 = 179.90, p < .01$)					($\chi^2 = 46.756, p = .045$)				
Average grade										
A (10)	2.4%	7.3%	20.0%	11.2%	76.0%	9.3%	22.2%	20.1%	27.0%	19.4%
A-/B+ (9)	23.1%	19.5%	17.3%	20.8%	3.9%	69.4%	61.7%	50.7%	49.4%	48.4%
B (8)	35.5%	34.0%	29.2%	32.5%	8.6%	16.0%	8.3%	20.1%	16.4%	22.6%
C (7) or lower	39.0%	39.2%	33.5%	35.6%	11.6%	5.3%	7.8%	9.0%	7.3%	9.7%
	($\chi^2 = 9.200, p < .01$)					($\chi^2 = 34.614, p = .022$)				

Table 4. Impact of satisficing on scale reliabilities, by study.

Scale	Scale Reliability				
	Non-Satisficers	Weak Satisficers	Moderate Satisficers	Strong Satisficers	Very Strong Satisficers
<i>Survey 1</i>					
Engagement (23 items)	0.841	0.839	0.844	0.873	0.896 ¹
Higher-level Thinking (5)	0.671	0.685	0.691	0.739	0.767 ¹
Development (17)	0.901	0.904	0.912	0.919	0.974 ¹
<i>Survey 2</i>					
Critical Thinking & Communication (9)	0.771	0.751	0.754	0.835	0.902 ¹
Arts & Humanities (6)	0.658	0.709	0.7	0.75	0.863 ¹
Independent, Confident Self (4)	0.69	0.584	0.624	0.717	0.82 ¹
Scientific/Quantitative Reasoning (3)	0.628	0.669	0.548	0.614	0.82 ¹
Interpersonal Skills (3)	0.745	0.608	0.566	0.618	0.776
Satisfaction with Major (15)	0.63	0.713	0.777	0.794	0.931 ¹

¹Difference between non-satisficers and very strong satisficers is statistically significant ($p < .01$) using Feldt's (1969) W statistic.

Table 5. Impact of satisficing on selected scale correlations, by study.

Scale	Scale Correlations				
	Non-Satisficers	Weak Satisficers	Moderate Satisficers	Strong Satisficers	Very Strong Satisficers
<i>Survey 1</i>					
Engagement -- Higher-Level Thinking	0.577	0.603	0.654	0.566	0.532
Engagement -- Development	0.486	0.482	0.453	0.493	0.658 ¹
Engagement -- Satisfaction	0.245	0.21	0.212	0.265	0.273
<i>Survey 2</i>					
Critical Thinking -- Arts & Humanities	0.408	0.342	0.476	0.495	0.671 ¹
Critical Thinking -- Scientific/Quantitative Reasoning	0.639	0.492	0.485	0.539	0.806 ¹
Critical Thinking -- Major Satisfaction	0.196	0.229	0.221	0.278	0.46 ²
Major Satisfaction -- Interpersonal Skills	-0.057	0.102	0.149	0.092	0.418 ¹

¹Difference between non-satisficers and very strong satisficers is significant at $p < .05$ using Fisher's z . ²Difference significant at $p < .1$