PACE
Policy Analysis for California Education
CORE-PACE RESEARCH PARTNERSHIP

WORKING PAPER

Assessing Survey Satisficing:

The Impact of Unmotivated Questionnaire Respondents on Data Quality

Education researchers use surveys widely. Yet, critics question respondents' ability to provide high-quality responses. As schools increasingly use student surveys to drive policymaking, respondents' (lack of) motivation to provide quality responses may threaten the wisdom of using surveys for data-based decision-making. To better understand student satisficing (sub-optimal responding on surveys) and its impact on data quality, we examined the pervasiveness and impact of this practice on a large-scale social-emotional learning survey administered to 409,721 students in grades 2-12. Findings indicated that despite the prevalence of satisficing in our sample, its impact on data quality appeared more modest than anticipated. We conclude by providing an accessible approach for defining and calculating satisficing for researchers, practitioners, and policymakers working with large-scale datasets.

Christine Calderon Vriesema
University of Wisconsin – Eau Claire

Hunter Gehlbach
Johns Hopkins University

VERSION: October 2019

Acknowledgements

This paper was produced as part of the CORE-PACE Research Partnership, which is focused on producing research that informs continuous improvement in the CORE districts (Fresno, Garden Grove, Long Beach, Los Angeles, Oakland, Sacramento City, San Francisco, and Santa Ana unified school districts) and policy and practice in California and beyond. We thank the CORE district administrators and leaders for their support and willingness to share their insights throughout the project. We also thank Collin Hitt, Joseph Robinson-Cimpian, Scott Barge, Nan Mu, Katherine Cornwall, Claire Chuter, Joe Witte, and Hans Fricke for reviewing and providing thoughtful feedback on earlier versions of this paper. PACE working papers are circulated for discussion and comment purposes and have not undergone the peer-review process that accompanies official PACE publications.

Introduction

Social scientists have long maintained a love-hate relationship with surveys. Educational researchers are no different. On the one hand, scholars love the capacity of surveys to uncover respondents' values, perceptions, and attitudes cheaply and at scale (Gehlbach, 2015; Gilbert, 2006; West, Buckley, Krachman, & Bookman, 2017). The intrinsic flexibility of surveys allows for respondents to report on: themselves (i.e., self-report measures), other individuals, or their perceptions of a whole class or community. Administration is typically efficient and straightforward—special training on how to complete a survey is rarely required (Duckworth & Yeager, 2015).

On the other hand, skeptics have leveled a broad array of critiques that question the value of survey data for researchers, practitioners, and policymakers. One broad set of criticisms questions whether respondents have the cognitive capacity for introspection that would be required to provide high-quality answers. For example, Nisbett and Wilson (1977) provided multiple illustrations of people's erroneous attempts to divine the real reasons behind their choices. Others have asked respondents how supportive they are of policies that do not exist–thus, showing how respondents may voice opinions that they could not possibly have (Bishop, Oldendick, Tuchfarber, & Bennett, 1980).

A second challenge arises from scholars who acknowledge that people might know their own attitudes, but worry that numerous forces conspire to inhibit respondents accurately reporting these attitudes back to the survey designer (Duckworth & Yeager, 2015). These forces include phenomena such as acquiescence bias, social desirability, floor/ceiling effects, biased question wording, response order effects, and so forth (e.g., Krosnick, 1999).

Yet, survey designers can relatively easily delimit surveys to asking about topics that respondents might reasonably have opinions on. Furthermore, they can design their surveys in line with many of the best practices that survey researchers have developed (Gehlbach & Artino, 2018). So, while these two potential problems with survey research as a methodology are real and need to be taken seriously, they are rarely insurmountable.

However, a third, potentially more challenging concern pertains to respondent motivation: sometimes participants do not take the survey seriously. In its most extreme form, some may become "mischievous responders" (Robinson-Cimpian, 2014) who are actively motivated to give false answers, perhaps out of an attempt to be funny. Krosnick (1991) describes milder, and potentially more prevalent forms of what he calls "satisficing", where respondents fail to put forth their best efforts. This motivation problem is sufficiently common that some researchers have even used effort (or lack thereof) on questionnaires as a performance task that they treat as a measure of conscientiousness (Hitt, Trivitt, & Cheng, 2016; Zamarro, Cheng, Shakeel, & Hitt, 2018).

As more and more schools are eager to use survey measures to inform policy decisions, this motivation problem may present the gravest concern yet. If a survey respondent wants to skip items, quit early, or speed through the survey by giving the same answer each time, survey designers can do little to prevent it. This problem highlights complementary needs for educators pursuing data-driven decision-making: to understand the pervasiveness of survey satisficing in large-scale student data and to determine the extent to which the satisficing impacts data quality. We address both needs by investigating satisficing in a regularly administered, large-scale survey of elementary and secondary students' social-emotional learning in California. This paper outlines straightforward strategies for detecting, assessing, and accounting for satisficing in large-scale student data. Hopefully, by arming researchers, practitioners, and policymakers with this information, they will develop wiser policies and interventions that reduce respondent satisficing.

Satisficing

Traditionally, the concept of satisficing refers to a cognitive heuristic in which people engage in sub-optimal decision-making strategies in order to conserve mental energy (Simon, 1957; Simon & Stedry, 1969). For example, rather than searching for an 'optimal' solution, some decision-makers select the first satisfactory alternative that seems 'close enough', thereby saving both time and effort. Within the context of survey design, survey scholars have adapted this concept to explain respondents' sub-optimal behaviors (Krosnick, 1991; Tourangeau, 1984). In survey research, satisficing can include selecting the first reasonable response option, agreeing with all the statements presented to the respondent, selecting the same option repeatedly in a straight-line across multiple items, and consistently selecting the "don't know" or "not applicable" responses (Barge & Gehlbach, 2012; Krosnick, 1991).

Although some survey researchers report potentially problematic respondent behavior in their work, few include systematic reports of survey satisficing. In one paper, Barge and Gehlbach (2012) examined the pervasiveness of survey satisficing and its effects on the reliability of and associations between scales for two surveys administered to college students. The authors found that the majority of students engaged in at least one form of satisficing (61% and 81% of students across the two surveys). This satisficing resulted in artificially inflated scale reliabilities and associations between scales. The pervasiveness of these practices and implications for data interpretation underscore the need to explore survey satisficing and its potential consequences in greater detail, especially for younger age-groups. This knowledge is particularly important now as large-scale data are increasingly used to guide decisions for policy and practice (Marsh, McKibben, Hough, Hall, Allbright, Matewos, & Siqueira, 2018).

Strategies for detecting satisficing include a range of methods that vary in complexity (Barge & Gehlbach, 2012; Steedle, Hong, & Cheng, 2019). Presumably researchers in all contexts (e.g., school districts, university settings, etc.) would benefit from considering the effects of respondent satisficing on their results. Ideally, any set of procedures to address satisficing should be accessible to as broad an audience as possible. Toward this end, we focus on three respondent behaviors that can be assessed within almost all survey data. The three response patterns include early termination–when respondents fail to complete the full survey; non-response, or missed items; and straight-line responding–when respondents select the same response option for at least ten items in a row. This definition of straight-line responding reflects prior work in this area (Barge & Gehlbach, 2012). Furthermore, this criterion fits the context of the present study: due to the substance of the items and the locations of reverse-scored items in the survey, the likelihood of 10 identical responses in a row being veridical seemed vanishingly small.

For this study, we operationalized satisficing as a respondent engaging in at least one of the three sub-optimal response patterns. Although other approaches exist (e.g., Robinson-Cimpian, 2014; Steedle, Hong, & Cheng, 2019), we focused on straightforward, accessible strategies for systematically defining, calculating, and reporting satisficing in large-scale student survey data in K–12 settings. By doing so, we hoped that these simple steps might be widely adopted by as many users of survey data as possible.

Research Questions and Hypotheses

To boost the transparency and credibility of our findings, we pre-registered a set of hypotheses (see https://osf.io/36zqk/) according to recommended practices (Gehlbach & Robinson, 2018). Specifically, we wanted to know (a) to what extent students engaged in survey satisficing; (b) which form of satisficing posed the largest threat to survey data; (c) which response option students were most likely to select when straight-lining in order to better discern how this strategy might impact students' mean scores on the survey scales, and (d) which students were most likely to satisfice.

Informed by our pilot results, we tested the following pre-specified hypotheses:

1. At least 10% of the total sample will engage in some form of satisficing.
2. Of the three types of satisficing examined, straight-lining will impact the greatest number of total survey items.
3. Straight-lining will impact the quality of the data. Specifically:
    a. Participants who straight-line will select the most extreme response option on the right-hand side of the scale the majority of the time.
    b. After accounting for reverse-scored items, straight-lining will significantly impact the mean scores.
4. Male students will satisfice more frequently than female students.

Method

Sample

This study examined secondary data collected by Policy Analysis for California Education (PACE). The dataset included student responses to a social-emotional learning (SEL) survey administered during the 2014–15 and 2015–16 school years. We used the 2014–15 school year data to conduct the exploratory, pilot analyses that guided our pre-registered hypotheses. The confirmatory, pre-registered analyses for this paper used data from the 2015–16 school year.

The pre-registered sample (*N* = 409,721) drew students from five California school districts. While two students in the sample were in Grade 2, most students ranged from grades 3 through 12. The sample included 146,126 elementary school students; 125,747 middle school students; and 137,838 high school students. For a complete description of student demographics, please see Table 1.

Table 1. Student Demographics

Student Characteristic	*N*	% of Sample
Gender		
Female	203,078	70.75%
Race/Ethnicity		
African American	35,256	8.60%
Asian	35,494	8.66%
Filipino	11,391	2.78%
Hispanic/Latino	289,862	70.75%
Native American	20,309	4.96%
Pacific Islander	3,312	0.81%
White	271,057	66.16%
Flagged District Designations		
Qualified for Free/Reduced Lunch	314,175	76.68%
Parents without High School Diplomas	95,788	23.38%
English Language Learner	70,118	17.11%
Homeless	10,303	2.51%
Student with Disability	45,977	11.22%
Suspension	5,417	1.32%
In-School Suspension	1,484	0.36%

Note: The percentages in each category may not equal 100 because the district data listed multiple designations for each student. For example, districts listed more than one race/ethnicity for 61.48% of the students in the sample.

Measures

Our analyses relied on students' responses to a 25-item social-emotional learning survey. The survey consisted of four scales measuring growth mindset (*n* = 4 items), regulation (*n* = 9 items), self-efficacy (*n* = 4 items), and social awareness (*n* = 8 items). Example items for each scale included the following: "My intelligence is something that I can't change very much," "I got my work done right away instead of waiting until the last minute," "I can do well on all my tests, even when they're difficult," and "How carefully did you listen to other people's points of view?" respectively. All items had five response options. We present reliability estimates in Table 2 and the complete scales and response options in Appendix A.

Table 2. Descriptive Statistics for the Complete and High-Fidelity Samples

	Complete Sample		High-Fidelity Sample		Feldt's *W*
Scale	α	*M* (*SD*)	α	*M* (*SD*)	
Growth Mindset	.72	3.76 (0.98)	.71	3.78 (0.95)	1.05
Regulation	.85	4.06 (0.68)	.83	4.05 (0.67)	1.08
Self-Efficacy	.87	3.53 (1.00)	.87	3.49 (0.98)	1.06
Social Awareness	.81	3.75 (0.71)	.80	3.73 (0.68)	1.10

Note: Feldt's *W* reflects the comparison between alpha coefficients for the complete and high-fidelity samples.

Procedures

To determine whether participants satisficed, we ascertained whether respondents engaged in at least one of three response patterns: early termination, nonresponse, and straight-line responding. This section summarizes the procedures used to calculate each response pattern; Appendix B provides in-depth descriptions of these calculations for practitioners, researchers, and policymakers interested in replicating these strategies.

For each form of satisficing, we determined whether respondents engaged in the specific response strategy or not (coded as 1 or 0, respectively). We operationalized early termination as ending the survey prior to completing the last survey item (i.e., item 25); participants were assigned a "1" if they ended early and a "0" if they did not. Non-response was operationalized as the number of items participants skipped in their survey. However, to avoid double-counting non-responders and early terminators, we first took into consideration how far each student got in the survey. We then calculated how many items students missed out of the total number of items they completed. Students who skipped at least one item were assigned a "1"; those who did not skip items were assigned a "0". To identify straight-line responding, we analyzed the standard deviation for each sequential set of 10 items across the complete survey (e.g., items 1–10, 2–11, 3–12, etc.). Standard deviations of zero for a given set indicated that the student selected the same response option for each of the 10 items. Thus, across the 16 possible intervals (i.e., the 16 sets of 10 sequential items), students received a "1" if they straight-lined at least one time. Finally, we determined overall satisficing—that is, whether a student satisficed at any point during the survey—by taking the sum of the three coded values yielded by these calculations; values greater than zero indicated that a student satisficed at some point during the survey.

Pre-registered Results

Hypothesis 1: Overall Satisficing

We tested our first hypothesis that at least 10% of the sample would engage in survey satisficing by dividing the number of students who satisficed by the total number of participants. Our data were congruent with this hypothesis with 30.36% of students engaging in at least one form of satisficing. The satisficing included 3.73% early termination, 24.99% nonresponse, and 5.38% straight-line responding. Some students engaged in multiple forms of satisficing (3.26% engaged in two forms and 0.14% engaged in all three).

Hypothesis 2: Survey Impact

We hypothesized that out of the three types of satisficing, straight-line responding would impact the greatest number of total survey items. In contrast to non-response and early termination, which minimally impact a single item, straight-line responding even one time implicates a minimum of 10 items, by definition. The results supported our pre-specified hypothesis in that the students who straight-lined engaged in this strategy for a mean of 3.90 intervals (each interval represents a set of 10 items; *SD* = 4.04). This average corresponds to selecting the same response option almost 13 items in a row. In comparison, average nonresponse corresponded to 1.77 skipped items, and early termination resulted in students ending an average of 3.52 items early.

Hypothesis 3a and 3b: Straight-Line Responding

We tested Hypothesis 3a, that participants who straight-lined would select the most extreme response option on the right-hand side of the scale a majority of the time, by obtaining the frequency distributions for each straight-lining interval (i.e., how often straight-line responding occurred for the first, second, third, etc. response option). We then calculated the percentage of straight-line responding that occurred using the response option on the far right-hand side for each interval. Participants selected this response option an average of 46.02% of the time across the 16 intervals–short of the 51% of participants we had predicted. The second most frequently selected option was the middle option (*M* = 29.97%).

To examine whether this response pattern impacted students' mean scores for the four scales, we conducted a series of two-sample *t*-tests for each of the four scales. We compared the complete sample to the high-fidelity responders (i.e., the sample after excluding respondents who straight-lined).[1]

The results supported Hypothesis 3b for each of the four scales. The complete sample had higher mean scores for the self-regulation (*t*(796909) = 9.68, *p* < .001, 99% CI: 0.01, 0.02; *Cohen's d* = .02), self-efficacy (*t*(794575) = 16.19, *p* < .001, 99% CI: 0.03, 0.04; *Cohen's d* = .04), and social awareness (*t*(795008) = 14.93, *p* < .001, 99% CI: 0.02, 0.03; *Cohen's d* = .03) scales than the high-fidelity sample. The same pattern emerged for the growth mindset scale; however, the items for this scale were reverse scored. Students who straight-lined on the far right-hand side of the scale (i.e., selecting response option 5) endorsed the conceptual opposite of growth mindset. Thus, after accounting for the reverse-scored growth mindset items, the mean growth mindset score for the complete sample was lower than the mean for the high-fidelity sample (*t*(794700) = -6.51, *p* < .001, 99% CI: -0.02, -0.01; *Cohen's d* = -.01). Please see Table 2 for the descriptive statistics for the two samples across each of the four scales.

Hypothesis 4: Identifying Satisficers

To test our hypothesis that male students would be more likely to engage in satisficing than female students, we conducted a logistic regression. The results supported our hypothesis, with male students being more likely to satisfice than female students (*B* = .15, *SE* = .01, odds ratio = 1.16, 99% CI = 1.14, 1.18).

Exploratory Results

Overall, our results showed the pervasiveness of student satisficing in our sample, particularly for male students compared to female students. The findings also highlighted the importance of identifying straight-line responding. This specific response pattern implicated the greatest number of survey items and significantly impacted students' mean scores on the four scales. Although these results supported most of our pre-registered hypotheses, important questions remained. In this exploratory results section, we first identified other student characteristics in addition to student gender that were associated with student satisficing. Second, given the influence of student straight-lining, we examined other outcomes associated with this response pattern in order to develop more refined recommendations for future research.

[1] We used 99% confidence intervals to evaluate our tests. We selected 99% confidence intervals in order to account for our five total hypotheses (i.e., the four hypotheses in 3b and Hypothesis 4). This corresponds to a critical *p*-value of .01.

Student Characteristics

To examine our first exploratory question, we fit a logistic regression model to examine whether other student characteristics also predicted survey satisficing. In addition to gender, we included race, grade, English Language Learner status, student with a disability status, free or reduced price lunch qualification, and suspensions. Results indicated that students were more likely to satisfice if they were in younger grades, English Language Learners, students with disabilities, students of color, male, and qualified for free or reduced price lunch. The number of suspensions did not predict student satisficing. Please see Table 3 for the complete list of district-reported student characteristics and the relevant statistical output.

Table 3. Student Characteristics that Predict Likelihood of Satisficing

		95% CI for Odds Ratio		
Satisficing	*B* (*SE*)	Lower	Odds Ratio	Upper
Intercept	-.27 (0.02)***			
Grade Level	-.13 (0.00)***	0.87	0.87	0.88
Male	.11 (0.01)***	1.10	1.12	1.13
Students of Color	.06 (0.01)***	1.03	1.06	1.09
English Language Learner	.20 (0.01)***	1.19	1.22	1.24
Student with Disability	.37 (0.01)***	1.42	1.45	1.48
Free/Reduced Lunch	.08 (0.01)***	1.06	1.08	1.10
Suspension	-.01 (0.02)	0.97	0.99	1.02

Note: *** *p* < .001. Grade level ranged from grades 2–12 and number of suspensions from 0–18. All other variables were dichotomous.

Impact of Straight-line Responding

Given the impact of straight-line responding on students' total completed survey items and students' mean scores for the four scales, we pursued several follow-up questions regarding this response pattern in particular. These exploratory research questions examined (a) how straight-lining impacted mean differences between specific student groups, (b) the impact of straight-line responding on the reliability of and correlations between scales, and (c) where students were most likely to satisfice in the survey.

First, we explored mean differences between student subgroups due to the interest in identifying potential gaps in student outcomes (e.g., the achievement gap). We focused on gender differences to follow up on our pre-registered finding that male students satisficed more often than their female counterparts. In this section, we specifically examined gender differences before and after removing student satisficers (i.e., the complete sample and the high-fidelity sample, respectively). Results indicated that mean scores for either group changed between 0.01 to 0.02 points after removing the students who satisficed (e.g., mean self-regulation scores for female students fell from 4.16 to 4.14). However, even though the mean scores changed, the magnitude of differences between female and male students remained consistent regardless of whether analyses were based on the complete sample or the high-fidelity sample. Female students reported higher self-regulation (*Cohen's d* = 0.28 for complete, 0.27 for high-fidelity), growth mindset (*Cohen's d* = 0.04 for complete, 0.03 for high-fidelity), and social awareness (*Cohen's d* = .22 for complete, 0.22 for high-fidelity) than male students. In contrast, male students reported higher self-efficacy than female students (*Cohen's d* = -0.08 for complete, -0.10 for high-fidelity).

Second, we compared alpha coefficients by using Feldt's (1969) *W* statistic. As Table 2 shows, the alpha coefficients for growth mindset, regulation, self-efficacy, and social awareness were between .01 and .02 higher for the complete sample as compared to the high-fidelity sample; these findings correspond to a *p*-value of less than .001. Please see Table 2 for Feldt's *W* output.

Third, we used Fisher's *z* to compare the correlation coefficients between the complete sample and the high-fidelity sample. Correlations for growth mindset with regulation (*z* = -12.65), self-efficacy (*z* = -13.23), and social awareness (*z* = -5.12) were higher for the complete sample than the high-fidelity sample. The same pattern emerged when examining the correlations for regulation with self-efficacy (*z* = 13.20) and social awareness (*z* = 13.16), as well as the correlation between self-efficacy and social awareness (*z* = 21.80). All correlations were significant at *p* < .001. See Table 4.

Table 4. Correlations Between the Four Survey Scales

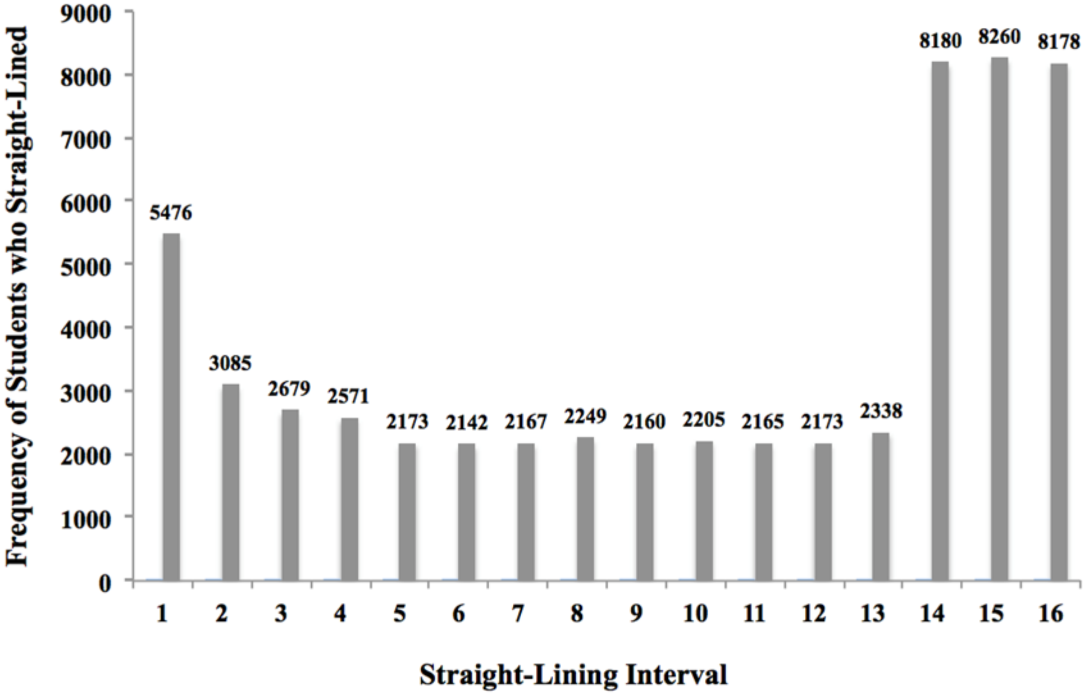
Scale	1.	2.	3.	4.
1. Growth Mindset	—	**0.23**	**0.28**	**0.13**
2. Regulation	0.21	—	**0.44**	**0.51**
3. Self-Efficacy	0.25	0.46	—	**0.43**
4. Social Awareness	0.12	0.53	0.47	—

Note: **Bold** correlations are for the non-straight-lining sample. Non-bolded correlations are for the complete sample.

Pattern of Satisficing

Thus, in our data, straight-line responding impacted mean scores, scale reliability, and correlation coefficients. This finding raises the question of where, within the survey, students straight-lined most frequently. We specifically examined this response pattern for all the students who otherwise completed the full survey (i.e., we included the students who did not skip items or end the survey early). We found that student straight-lining (a) decreased after the first interval, (b) remained fairly consistent for the next 13 intervals (i.e., sets of 10 items), but (c) increased during the last three intervals of the survey. See Figure 1.

Figure 1. Pattern for Straight-Line Responding



Discussion

As schools increasingly use student survey measures to inform practice and policy decisions, the need to explore the pervasiveness of students' survey satisficing grows more critical. We took a deliberately simple approach to defining and calculating satisficing so that our approach might be easily replicated, even in the absence of highly trained statisticians. Toward this end, we identified three response patterns—early termination, non-response, and straight-line responding—and provided a set of accessible strategies for detecting respondent satisficing. Moreover, we pre-registered five hypotheses to examine the pervasiveness and impact of survey satisficing within a large-scale student survey. In this section, we briefly discuss our findings and then provide a set of recommendations around respondent satisficing for researchers, practitioners, and policymakers.

Total Satisficing

Findings indicated that students satisficed extensively in our large-scale dataset. Overall, a little more than 30% of the sample engaged in at least one form of satisficing. Given that survey satisficing reflects a lack of respondent motivation, however, it is important for researchers to consider how they define the individual response patterns. We took an inclusive approach to our definitions. In particular, respondents qualified as non-responders as long as they missed one item. It is possible, though, that this inclusive definition misrepresented student metadata (Soland, Zamarro, Cheng, & Hitt, 2019), inadvertently categorizing some motivated students who may have skipped items accidentally rather than deliberately as satisficers. Researchers examining satisficing in their own data will need to determine meaningful definitions for their specific contexts.

Impact on the Survey

Of all three response patterns, student straight-line responding impacted the greatest number of total survey items. Moreover, because of the potential for straight-line responding to impact students' mean scores on the survey, we examined which response option students selected when using this strategy and how this impacted the scores on the four survey scales.

Focusing on straight-line responding, findings indicated that the students who straight-lined selected the response option on the far right-hand side almost half the time (*M* = 46.02%). We are confident that students are not accurately reporting their attitudes because the survey included a set of reverse-scored items measuring growth mindset. The right-hand response option therefore signaled a fixed mindset—that is, the conceptual opposite of growth mindset. The students who reported the lowest growth mindset also likely endorsed the highest self-efficacy and regulation. These findings would be incongruous with the motivation research linking stronger growth mindsets with higher self-efficacy (Dweck & Master, 2009).

When examining the impact of straight-line responding on mean scores, we found that mean scores differed significantly between the complete sample and the high-fidelity sample (i.e., the sample after excluding straight-liners). Yet, the relatively modest effect sizes suggest that, while significant, the differences between samples may not necessarily represent a substantial threat to interpretations of the findings. For example, the *Cohen's d* coefficients ranged from .01 to .04; these coefficients fall below the .20 cutoff typically reserved for 'small' effect sizes (Cohen, 1988). Of course, the magnitude of effect sizes ranges across research contexts—what may be a small effect size in one domain may represent a meaningful difference in others. Moreover, some researchers argue that effect size cutoffs are relatively arbitrary and should instead be interpreted in terms of the consequences that the effects could cause (Funder & Ozer, 2019). Local contexts can therefore help guide when the differences between the full sample and high-fidelity sample are meaningful.

The exploratory analyses investigating scale reliability and correlations between survey scales also yielded relatively small differences between the two samples. In spite of these modest differences overall, we explored whether removing satisficers might have a greater impact on student sub-group analyses. Similar to the overall analyses, however, results indicated that the magnitude of differences between female and male students remained consistent even after removing student satisficers. Together, the findings suggest that while researchers need to be aware of how straight-lining impacts data quality in their respective samples, the response pattern may not always threaten the integrity of the results.

Respondent Characteristics

Beyond investigating the impact of satisficing on data quality, we also examined whether satisficing might change the nature of the sample in systematic ways. Addressing the problem of respondent motivation by removing satisficers from the sample could lead to unrepresentative samples if certain groups satisfice at higher rates than others. For example, in the present sample, findings supported our pre-registered hypothesis that male students were more likely to satisfice than their female counterparts. Our exploratory analyses also identified characteristics such as race/ethnicity, language status, and disability status as other factors associated with satisficing. However, while significant, the effects of these characteristics also tended to be modest. As a result, the policy decisions stemming from such research might be misguided for certain groups. On the other hand, leaving satisficers included in analyses could also lead to unrepresentative findings given that satisficing—at least in this sample—occurs at different rates for different groups of students. As noted above, researchers will need to determine to what extent satisficing impacts their own data quality in order to ascertain whether it is necessary to remove satisficers prior to presenting their findings.

Recommendations for Researchers, Practitioners, and Policymakers

Based on this study, we recommend the following five guidelines. First, researchers, practitioners, and policymakers will need to determine meaningful definitions of satisficing that make sense within the context of their surveys. In our study, we took an inclusive approach to satisficing in order to determine the most extensive effects of students' sub-optimal response patterns. However, for some districts or policymaking settings, taking a more conservative approach might be more appropriate (e.g., defining non-response as four missed items rather than one missed item). Fortunately, testing different definitions of satisficing and examining the repercussions is relatively low cost—merely the time taken to conduct additional analyses. As data analysts further explore the impact of satisficing, we recommend testing various definitions to see what is most sensible for a given context.

Second, we recommend that researchers, practitioners, and policymakers examine their data with and without straight-line responders in order to evaluate how much this response pattern affects interpretation of the findings. Our results indicated that straight-lining did not substantially change interpretation of the findings; however, given the context-dependent nature of education, these results might look different in a different setting.

Third, we recommend against excluding all data from every student who satisfices. Instead, researchers, practitioners, and policymakers may benefit more from removing only the flawed data. Specifically, because the strategies of skipping items and early termination result in missing data, data analysts really need to be concerned only with straight-line responders. Therefore, while we recommend that analysts first confirm that excluding straight-line responders does not markedly change the student demographics in their samples, removing flawed data may support those in applied settings to use only quality data to inform decision-making. Removing these flawed data (i.e., the data for straight-line responders) will also help to ensure that analysts are not throwing quality data away along with the potentially compromised data (i.e., flawed data that could potentially alter the means for the four scales).

Fourth, including reverse-scored items in a survey may seem like an effective strategy for detecting straight-line responders. However, we caution against using this tactic. Reverse-scored items often do not fall on a continuum, reduce scale reliability, and are difficult for participants to answer (Benson & Hocevar, 1985; Gehlbach & Brinkworth, 2011; Swain, Weathers, & Niedrich, 2008). Instead, survey designers can attempt to mitigate straight-line responding by interspersing items from different constructs (Gehlbach & Barge, 2012) and ensuring that response options are item-specific (Gehlbach & Brinkworth, 2011).

Lastly, because students' motivation to respond to survey items carefully is malleable, we urge those using survey research to cultivate buy-in from students prior to administering the survey instruments. Using evidence-based strategies (e.g., Dillman, Smyth, & Christian, 2014) to enhance respondent motivation early in the survey process may reduce some of the satisficing behaviors utilized by students.

Conclusion

Critiques of survey data abound. These criticisms question respondents' ability to understand their own attitudes, accurately report their attitudes, and engage with survey with sufficient motivation. This third critique—low respondent motivation—may present the largest potential threat to data interpretation given the lack of researcher control over this type of respondent behavior. However, our findings indicate that despite the prevalence of student satisficing in our sample, the impact of this practice on data quality appeared surprisingly small. Because of the context-specific nature of education, we urge others to similarly determine the prevalence and impact of survey satisficing in their own datasets. Through a collective effort, we can learn how robust survey findings are to students who engage in satisficing behaviors. To support researchers, practitioners, and policymakers pursuing this important task, we provided an accessible foundation for defining and calculating student satisficing in large-scale datasets. We hope these strategies ultimately facilitate those individuals who are trying to help schools to make better data-driven decisions.

References

Barge, S., & Gehlbach, H. (2012). Using the theory of satsificing to evaluate the quality of survey data. *Research in Higher Education, 53*(2), 182–200.

Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitudes scales for elementary school children. *Journal of Educational Measurement, 22*(3), 231–240. doi:10.1111/j.1745-3984.1985.tb01061.x

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1980). Pseudo-opinions on public affairs. *The Public Opinion Quarterly, 44*(2), 198–209.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: *The tailored design method* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher, 44*(4), 237–251. doi:10.3102/0013189x15584327

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), p. 156-168.

Gehlbach, H. (2015). Seven survey sins. *The Journal of Early Adolescence*, 35, 883–897. https://doi.org/10.1177/0272431615578276

Gehlbach, H., & Artino, A. R. (2018). The survey checklist (manifesto). *Academic Medicine: Journal of The Association of American Medical Colleges, 93*(3), 360–366. doi:10.1097/ACM.0000000000002083

Gehlbach, H., & Barge, S. (2012). Anchoring and adjusting in questionnaire responses. *Basic and Applied Social Psychology, 34*(5), 417–433. doi:10.1080/01973533.2012.711691

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology, 15*(4), 380–387. doi.org/10.1037/a0025704

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness, 11*(2), 296–315. doi:10.1080/19345747.2017.1387950

Gilbert, D. T. (2006). *Stumbling on happiness* (1st ed.). New York: Alfred A. Knopf.

Hitt, C., Trivitt, J., & Cheng, A. (2016). When you say nothing at all: The predictive power of student effort on surveys. *Economics of Education Review, 52*, 105–119. https://doi.org/10.1016/j.econedurev.2016.02.001

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236.

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

Marsh, J. A., McKibben, S., Hough, H. J., Allbright, T. N., Matewos, A. M., & Siqueira, C. (2018). *Enacting social-emotional learning: Practices and supports employed in CORE districts and schools.* Policy Analysis for California Education, PACE.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259.

Nunnally, J. C. (1968). *Psychometric theory* (2nd edition). New York: McGraw-Hill.

Robinson-Cimpian, J. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educational Researcher, 43*(4), 171–185.

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4*(1), 61–79.

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

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

Soland, J., Zamarro, G., Cheng, A., & Hitt, C. (2019). Identifying naturally occurring direct assessments of social-emotional competencies: The promise and limitations of survey and assessment disengagement metadata. *Educational Researcher, 48*(7), 466–478.

Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social-emotional learning competencies. *Educational Measurement: Issues and Practice.* doi:doi.org/10.1111/emip.12256

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research, 45*(1), 116–131. doi:10.1509/jmkr.45.1.116

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.

West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2017). Development and implementation of student social-emotional surveys in the CORE Districts. *Journal of Applied Developmental Psychology*, (in press).

Zamarro, G., Cheng, A., Shakeel, M. D., & Hitt, C. (2018). Comparing and validating measures of non-cognitive traits: Performance task measures and self-reports from a nationally representative internet panel. *Journal of Behavioral and Experimental Economics, 72,* 51–60. doi:10.1509/jmkr.45.1.116

Appendix A

The following table includes the 25 survey items in the order they were presented to students. For the reader's convenience, scale titles are displayed in the table.

Item	Response Options				
Scale: Regulation					
1. I came to class prepared.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
2. I remembered and followed directions.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
3. I got my work done right away instead of waiting until the last minute.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
4. I paid attention, even when there were distractions.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
5. I worked independently with focus.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
6. I stayed calm even when others bothered or criticized me.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
7. I allowed others to speak without interruption.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
8. I was polite to adults and peers.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
9. I kept my temper in check.	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
Scale: Growth Mindset					
10. My intelligence is something that I can't change very much.	Not At All True	A Little True	Somewhat True	Mostly True	Completely True
11. Challenging myself won't make me any smarter.	Not At All True	A Little True	Somewhat True	Mostly True	Completely True
12. There are some things I am not capable of learning.	Not At All True	A Little True	Somewhat True	Mostly True	Completely True
13. If I am not naturally smart in a subject, I will never do well in it.	Not At All True	A Little True	Somewhat True	Mostly True	Completely True
Scale: Self-Efficacy					
14. I can earn an A in my classes.	Not At All Confident	A Little Confident	Somewhat Confident	Mostly Confident	Completely Confident
15. I can do well on all my tests, even when they're difficult.	Not At All Confident	A Little Confident	Somewhat Confident	Mostly Confident	Completely Confident
16. I can master the hardest topics in my classes.	Not At All Confident	A Little Confident	Somewhat Confident	Mostly Confident	Completely Confident
17. I can meet all the learning goals my teachers set.	Not At All Confident	A Little Confident	Somewhat Confident	Mostly Confident	Completely Confident

Scale: Social Awareness					
18. How carefully did you listen to other people's points of view?	Not Carefully At All	Slightly Carefully	Somewhat Carefully	Quite Carefully	Extremely Carefully
19. How much did you care about other people's feelings?	Did Not Care At All	Cared A Little Bit	Cared Somewhat	Cared Quite A Bit	Cared A Tremendous Amount
20. How often did you compliment others' accomplishments?	Almost Never	Once in a While	Sometimes	Often	Almost All the Time
21. How well did you get along with students who are different from you?	Did Not Get Along At All	Got Along A Little Bit	Got Along Somewhat	Got Along Pretty Well	Got Along Extremely Well
22. How clearly were you able to describe your feelings?	Not At All Clearly	Slightly Clearly	Somewhat Clearly	Quite Clearly	Extremely Clearly
23. When others disagreed with you how respectful were you of their views?	Not At All Respectful	Slightly Respectful	Somewhat Respectful	Quite Respectful	Extremely Respectful
24. To what extent were you able to stand up for yourself without putting others down?	Not At All	A Little Bit	Somewhat	Quite A Bit	A Tremendous Amount
25. To what extent were you able to disagree with others without starting an argument?	Not At All	A Little Bit	Somewhat	Quite A Bit	A Tremendous Amount

Appendix B

For all three forms of satisficing, we determined whether respondents engaged in the specific response pattern (coded as "1") or not (coded as "0"). We used these dichotomous variables in our analyses. However, the steps for calculating each response pattern also yield percentages. Thus, researchers, practitioners, and policymakers can decide to use one or both of these variable types in their own work depending on whether their research questions would benefit from dichotomous or continuous data.

Early Termination

First, we recoded participants' responses to each survey item. If a student provided any numerical response, the value was recoded as the relevant item number in the survey (e.g., the value for the first item was recoded to "1"; the value for the second item was recoded as "2", etc.). We then used Stata to identify the last non-missing number. Finally, we divided the last non-missing number by the total number of possible items to determine the percentage of the survey that students completed. Values totaling 100 were recoded "0" to signify that no early termination occurred; values less than 100 were recoded as "1" to signify that respondents ended the survey before completing the last item.

Nonresponse

We determined whether students skipped items at any point in the survey by using Stata to determine the total number of missed items across the full survey. However, we took into consideration whether students ended the survey early by subtracting the number of items missed due to early termination. We then calculated the percentage of missing data by dividing the number of missing items by the total number of items completed. By using the total number of items completed rather than 25 (i.e., the total number of survey items), we ensured that early terminators were not also automatically coded as non-responders. Values of zero were recoded as "0" to signify that there were no instances of nonresponse; values greater than zero were recoded as "1" to signal that nonresponse occurred.

Straight-line Responding

We determined whether students straight-lined at any point in the survey by analyzing sequential sets of 10 items for the complete survey (e.g., items 1–10, 2–11, 3–12, etc.). First, for each set of 10 consecutive items in each survey, we created a new variable indicating its standard deviation. If the standard deviation for a set of given items was zero, the value was recoded as "1", indicating that the student straight-lined; if the standard deviation was any non-zero number, that value was recoded as "0". Next, we determined the percentage of straight-lining that occurred by (a) determining the number of times a student straight-lined and then (b) dividing that number by the total number of possible intervals. Values of zero were coded as “0” to indicate that straight-lining did not occur; values greater than zero were coded as “1” to indicate that straight-lining did occur.

Satisficing

We determined whether participants satisficed by calculating the sum of the participants’ coded scores for the three response patterns. Values equal to zero were recoded as “0” signifying that no satisficing occurred. Values greater than zero were recoded as “1” to signal that a participant engaged in at least one form of satisficing.