Survey Research Methods (2020)
Vol. 14, No. 5, pp. 439-459
doi:10.18148/srm/2020.v14i5.7641

ISSN 1864-3361
http://www.surveymethods.org

Valid vs. Invalid Straightlining: The Complex Relationship Between Straightlining and Data Quality

Kevin Reuning
Miami University
Oxford OH

Eric Plutzer
Penn State University
University Park PA

Straightlining—the tendency to give the same response to a series of grouped questions—can be the result of satisficing respondents. As a result, many survey practitioners use straightlining as one, and sometimes the only, indicator of data quality. Respondents identified as straightliners are often removed from the data set on the assumption that their answers are meaningless. In this paper we show that these practices are based on a logical fallacy and demonstrate that in many common survey formats, the incidence of straightlining can be increased by improving the validity and the reliability of survey questions. We take initial steps in investigating the complexities and challenges of data analysis by providing a formal definition of *valid straightlining* and leverage that definition in a series of Monte Carlo simulations to better understand the conditions that give rise to valid straightlining. Although it remains for future work to distinguish valid from invalid straightliners, our formal definition of the concept and our simulation methods augment the tools survey analysts employ in assessing the prevalence of low effort respondents in survey data sets. The paper thereby takes initial steps toward sounder methods of classifying straightliners as optimizers or satisficers.

Keywords: satisficing; data quality; psychology of survey response

When survey respondents pay minimal attention to questions or make little effort to answer accurately (satisficing) data quality is compromised. One widely recognized symptom of satisficing is *straightlining* —the tendency of survey respondents to provide identical answers to consecutive questions. Of the many indicators of satisficing, straightlining is among the most frequently investigated because it is easy to identify and quantify (Kim, Dykema, Stevenson, Black, & Moberg, 2019).

The assumption that straightlining is a good measure of data quality is widely embraced. Fricker, Galesic, Tourangeau, and Yan (2005) use straightlining as one way to evaluate whether some modes of data collection yield superior data quality than others. Zhang and Conrad (2014) use straightlining as an indicator of poor response quality and use measures of straightlining to determine when survey speeding increases measurement error. Summarizing the consensus, Yan (2008, p. 3) observes that "straightlining is a form of measurement error and thus decreases data quality."

We argue that this characterization is overly broad, and empirically incorrect under many common conditions. Many practitioners understand this intuitively, but intuition is a poor guide for sound science. Therefore, we formalize this intuition by defining the concept of *valid straightlining.* Using this definition and Monte Carlo simulations, we estimate its prevalence in typical sets of survey questions and find that *under many common conditions, straightlining becomes more prevalent with (1) increasing validity and (2) increasing reliability.*

However, measurement quality is a means to more fundamental ends — obtaining unbiased parameter estimates that help answer important theoretical and policy questions. Thus we also assess the sensitivity of substantive findings to different straightlining scenarios by embedding our simulations within the data from the Health and Retirement Survey (2014). We show the removal of valid straightliners can lead to substantially biased estimates of coefficients intended to test theoretically informed hypotheses.

These results are important because many researchers use straightlining as a criterion for excluding respondents from analyses, based on the untested assumption that their removal improves overall data quality (e.g. Greszki, Meyer, & Schoen, 2014, p. 239; see also Bethlehem & Biffignandi, 2012, pp. 111–113). Indeed, at least one polling firm developed software to efficiently remove straightlining respondents (Schonlau & Toepoel, 2015). Major government surveys also engage in this practice (e.g., Centers for Disease Control 2010).

Our findings should give researchers pause when consid-

Contact information: Eric Plutzer, Department of Political Science, Penn State University, University Park, PA USA 16802 (E-mail: exp12@psu.edu)

ering the removal of respondents who give the same or similar answers to a sequence of questions. Although it remains for future work to distinguish valid from invalid straightliners, this paper augments the tools survey analysts employ in assessing the prevalence of low effort respondents in a dataset. We thereby advance our understanding of a very important, but "understudied" topic (Kim et al., 2019, p. 2)

1 Is Straightlining a Valid Indicator of Data Quality?

Herzog and Bachman (1981) coined the term straightlining to refer to giving identical answers to an entire set of questions. Krosnick (1991) placed straightlining in the context of respondent satisficing—applying minimal thought and effort when answering survey questions. From Krosnick's perspective, straightlining can result from completely thoughtless responses (deciding to answer every question with the "agree" or "disagree" response). It can also result from respondents skimming quickly and selecting a reasonable answer on a scale that more or less applies to all questions in a set (weak satisficing). And while Krosnick expected straightlining to be especially common when respondents rated attitude objects on a common scale, satisficing can lead to straightlining on many kinds of survey questions, including self-reports of behavior and assessments of social conditions, as well as attitudes.

Straightlining is more common in self-administered modes than in live interviewer surveys, and when questions are presented in a visual grid format but occurs even when grid-style questions are presented one at a time (Kim et al., 2019). Respondents who speed though surveys have been found to have higher rates of straightlining (Zhang & Conrad, 2014).

Straightlining is frequently used to measure data quality at the level of individual respondents, but also as an aggregate measure applied to question batteries and entire studies. Straightlining has been tied specifically to inattentiveness (Greszki et al., 2014) and is a particular concern in surveys of teens and young adults, who may not take surveys seriously (e.g. Cole, McCormick, & Gonyea, 2012; D. Cornell, Klein, Konold, & Huang, 2012). However, scholars have identified elevated levels of straightlining in virtually all types of surveys, across many populations, and on many topics.

Investigators who anticipate high levels of satisficing incorporate methods to identify low-effort respondents. Computer administered surveys can use time stamps to measure speeding. Others employ "trap questions"—instructional manipulation checks—and other devices as secondary indicators of random answers, inattentiveness and other threats to data quality. Indeed, some surveys ask respondents directly if they answered all questions truthfully (D. Cornell et al., 2012). Unfortunately, not all studies can easily incorporate such checks. Moreover, analysts of secondary data may only have the option to calculate measures of non-differentiation if they suspect high levels of satisficing.

1.1 Valid vs. Invalid Straightlining, and the Fallacy of Falsely Affirming the Consequent

Straightlining is one manifestation of satisficing. However, if satisficing frequently leads to straightlining, it does not logically follow that straightlining is a good indicator of data quality. While a sleepless night can lower test performance, it would be incorrect to view poor test performance as a good indicator of sleeplessness. This kind of inference would be logically sound if, *and only if*, satisficing is the only cause of straightlining— otherwise it reflects the *fallacy of falsely affirming the consequent.*

Indeed, the literature includes many instances when we expect respondents to give the same answer to a series of questions even when they are doing their best to understand the question, thoughtfully consider their answer, and carefully record their response. For example, teenagers who have never been victims of bullying should report "never" to a series of questions asking about victimization experiences. Likewise, fiscal conservatives can be expected to always answer "reduce federal spending" when asked about a series of policy domains such as education, national parks, transportation, and so on.

These examples provide the intuition for how the underlying distribution of the latent variable combines with attributes of survey questions to contribute to the frequency of valid straightlining. They illustrate what Schonlau and Toepoel (2015) refer to as "plausible" straightlining because non-differentiation is expected in the absence of satisficing behavior.

2 Valid Straightlining and the Psychology of Survey Response

Most survey methodologists understand that non-differentiation is expected or "plausible" under some conditions. One contribution of this paper is to formalize the notion of "plausible" straightlining as *valid straightlining,* and we begin with a definition:

Valid straightlining occurs when two conditions are satisfied: A respondent is motivated to carefully read/listen to questions and answer them truthfully, and provides identical (non-differentiated) responses to a sequence of questions.

A second contribution is to integrate the concept of valid straightlining into the "psychology of survey response" (or "cognitive aspects of survey methodology", CASM) model. CASM views the translation of latent variable values to survey responses as a multistage process (Tourangeau, Rips, & Rasinski, 2000a, 2000b). Each stage has important implications for the expected levels of valid and invalid straightlining.

2.1 Comprehension, recall and judgment

The process of answering survey questions has four stages: respondents (1) understand the question, (2) retrieve relevant cognitions from memory, (3) integrate/evaluate this information to arrive at a mental answer, and then (4) report this mental answer consistent with the format of the survey. *Satisficing survey respondents* can fail to execute any step (Krosnick, 1991): they might not understand the question if they read too quickly or do not attend to the speech of an interviewer; they may settle on the first cognition that comes to mind (Zaller et al., 1992); they can formulate a mental answer to the question without much thought or judgment; and they can record their answer haphazardly or randomly. *Optimizers*, in contrast, make some effort to arrive at an accurate mental answer and faithfully translate that answer to the most appropriate response.

We want to build a model of survey responses for survey optimizers—not satisficers—because by our definition only optimizers can engage in valid straightlining. To that end it is useful to combine the first three stages as a cognitive mechanism that translates a latent variable into a *mental answer: the rating, estimate, judgment or evaluation that a good faith, optimizing respondent arrives at before recording or reporting that answer*. With that simplification, we create of formal model with the following elements:

Latent variable: We assume a reflective measurement model, with a "true" latent variable such as a respondent's "true" ideological position, her "sincere and informed" policy preference, the "actual" number of times that he was bullied, or her family's "real" level of food insecurity.

The *mental answer* is the preliminary answer to the question posed in the questionnaire (the culmination of the first three steps in the CASM model). If the question stem asks "how often," the respondent may develop a mental answer in terms of loose verbal signifiers ("frequently", "a lot", "not too often") or more precise descriptions ("twice last week"). *We simplify by assuming that mental responses are measured on a continuous dimension*, and that this is the same dimension as the latent variable with the same mean and measurement scale. And for our purposes, we assume that even if two respondents use different verbal signifiers, their mental answers map onto the same latent scale.

Validity: Consistent with classical measurement theory, we define validity as the effect of the latent variable on the expected value of the mental answer (Bohrnstedt, 2010), parameterized as the factor loading measuring how well a survey question reflects a latent variable (e.g. Saris, Revilla, Krosnick, & Shaeffer, 2010). To keep our initial simulation simple, we save our examination of systematic bias for our third illustration.

Unreliability is defined as the variance of the random component of the mental answer. In our first example, we simplify by assuming that random components are drawn from a normal distribution—however, the model allows for any realistic probability distribution of random errors.

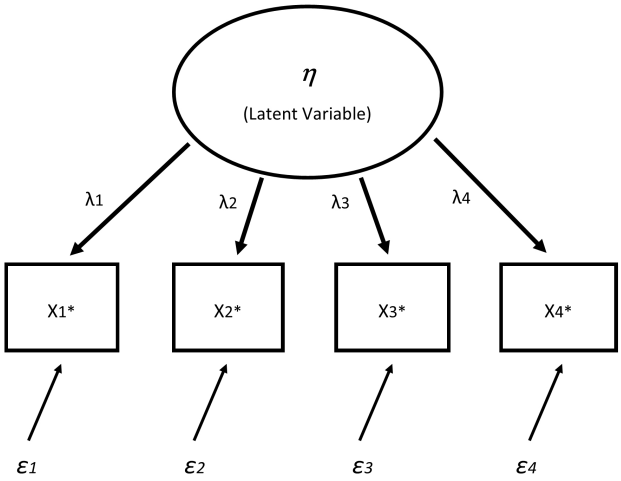

Figure 1. Reflective measurement model predicting mental answers (X_J^*) to four survey questions

2.2 Formalization

The mental answers X^* of respondent i to a set of J questions can be expressed as a reflective measurement model as illustrated in Figure 1 and algebraically as:

$$X_{Ji}^* = \lambda_J \eta_i + \epsilon_{Ji},$$

where η_i (Eta) is respondent i's true score on the latent variable, X_{Ji}^* ("X-star") is respondent i's mental answer to question J, with X^* assumed to be on the same scale as η, λ_J (Lambda) is the validity of the mental answer (X_J^*), as a measure of η_i, ε_{Ji} (Epsilon) is a random error for person i in formulating a mental answer to question J.

Several implications follow directly from the model. First, if the mental answers X^* are measured on the same scale as the latent variable, then the mental answers will be identical (non-differentiated) when $\lambda_1 = \lambda_2 = \lambda_3 = \ldots = \lambda_J$ and the variance of all $\varepsilon_J = 0$.

More generally, when all λs have the same sign, "mental straightlining" will be most frequent when reliability is equal to 1 ($\sigma_{\varepsilon,J} = 0$). That is, high reliability increases the incidence of valid straightlining because lower reliability will perturb the locations of mental answers, making it less likely they consistently align with the value of the latent variable. The counter-intuitive implication is that holding all else constant, straightlining is a symptom/indicator of both *high* data quality and a symptom of low effort respondents.

The second implication of this model is that when all items have equal validity (again, at the level of the mental answer and assuming scale equivalence and λs of the same sign), straightlining is maximized.

Validity—expressed by the magnitudes of the λ_J coefficients—has a somewhat different impact. Holding all else constant, low average validity will serve to attenuate relationships and mental answers will be biased toward the mean. For example, if the latent variable is distributed normally with approximately 99% of respondents lying between -3 and +3, then when $\lambda_J = 1$ and $\text{Var}(\varepsilon_J) = 0$, the mental answers will have the same range (-3 to +3). If validity is, say, 0.67, then mental answers will lie between -2 and +2. Thus, we can deduce that high validity can stretch mental answers across a wider range of the underlying dimension and low validity will lead to more answers lying close to the center of the distribution. Whether or not these distributional consequences lead to more or less straightlining depends, we will show, on the item difficulties—the thresholds that determine how mental answers map onto available answer choices.

However, as a general matter the higher the validity, the more straightlining when all question stems are worded in the same direction. To see this, consider the extreme case in which all $\lambda = 0$ and so validity is zero, and responses are solely determined by random draws from the distribution of ε. Valid straightlining, then, can only occur by chance. As validity increases, mental answers X^* get closer and closer to their corresponding value of the latent variable and valid straightlining becomes possible. We will show that this does not consistently hold when one or more questions are worded in the opposite direction.

In contrast, lower reliability perturbs the locations of mental answers, generally producing a wider distribution with a lower peak and fatter tails. Both low validity and low reliability have implications for how respondents map their answers on to the available response options in the survey.

2.3 Reporting or recording the mental answer

In the last stage of the CASM model, respondents translate their mental answers into recorded responses. We will assume that response categories are mutually exclusive, exhaustively cover the range of the latent dimension characterizing mental answers, and that optimizers make an effort to accurately record the answer that most closely aligns with their mental answer.

Under these assumptions, the semantic labels of a survey question create cut-points along the underlying dimension. In some cases, this is straightforward, as in responding to the question stem, "*How many days* in the last month were you verbally bullied during the school day?" A variety of response options are possible, including those phrased in terms of a specific number of days (e.g., "one to three days" or "ten days or more") and those with looser verbal signifiers (e.g., "never," "once in a while," "often," "every day"). Questionnaire design decisions, particularly about the wordings for the highest and lowest categories can raise or lower the number of valid straightliners. More optimizers will select "10 days or more" than "fifteen days or more" whenever the distribution of true scores extends to that end of the distribution. Similarly, a five-point rating scale whose last category is "*strongly* agree" might generate scores of "5" more frequently than one in which the last choice is "*completely* agree" since the latter is logically a special case of the former (Lieberman, Hancuch, & Buttermore, 2019).

This means that semantic labels create implicit cut points that interact with the shape of the underlying distribution to generate the data, thereby impacting the likelihood that any one optimizing respondent provides the same response to each question.

Formally, we can incorporate this into the final CASM stage as follows:

$$X_{Ji} = K \quad \text{if} \quad \kappa_{K-1} < X^*_{Ji} \leq \kappa_K \quad ,$$

where X^*_{Ji} ("X-star") is i's mental answer to question J, X_{Ji} is i's recorded answer (survey response) to question J, K is a vector of integers running from 1 to the number of answers offered, κ (Kappa) is the value of X^*_J that separates answers of K from answers of $K - 1$.

We now have a simple model for the formulation of mental answers to survey questions and the mapping of mental answers on to available survey response. In a sample composed only of optimizing respondents (zero satisficers), the number of valid straightliners and the response choices that reflect straightlining will therefore depend on:

1. The distribution of η, the latent variable
2. The number of questions in the grid or sequence (J)
3. The validities of the items (λ_J) defined in terms of mental answers
4. The standard deviation of the random error (σ_ε) defined in terms of mental answers
5. The values of thresholds (κ) that determine how mental answers map onto responses.

Below, we will also consider the number of items which are intended to be reverse scored.

3 Illustration #1—The Diener Satisfaction with Life Scale

Our first example is the Satisfaction with Life Scale. The paper introducing this scale (Diener, Emmons, Larsen, & Griffin, 1985) has been cited more than 20,000 times, and a version of it was employed in the Health and Retirement Survey as shown in Figure 2.

Because it was included in a self-administered pencil-paper component of the HRS, the conditions for satisficing are considerable. Without time stamping to identify speeders, analysts might be tempted to identify straightliners for possible removal. In the 2006 wave this would remove 1,198 respondents (16.4% of the sample).

The Diener Scale is typical of questionnaire sequences common in major social surveys. Grids or sequences with

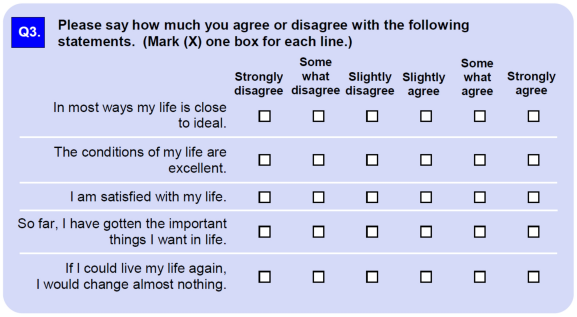

Figure 2. Satisfaction with Life Scale as implemented in the 2006 wave of the Health and Retirement Survey

ordinal response and lacking reverse worded question stems are regularly employed in the World Values Survey[1], International Social Survey Program[2], and the American National Election Study[3], to name just a few.

For this illustration we generated a simulated latent variable in a five-step process. We began with respondents' actual answers combined into an additive scale running from 5 to 30. Because the scale produced substantial floor and ceiling effects, we randomly added or subtracted up to five scale points to those heaped at the lowest and highest possible scores, extending the range of the variable to now run from 0 to 35. To smooth the distribution due to heaping at scores of ten and twenty, we added a spherical disturbance with standard deviation of 0.20. And we further smoothed the distribution by adding an additional spherical disturbance to all cases (sd = 0.20). We then standardized the variable to have a standard deviation of 1.0 and a mean of 0.60, to match the actual distribution of scores (slightly left skewed because most people report being satisfied). A full description how we constructed the latent variables is provided in Appendix C. Materials to replicate all tables and figures in this article can be found in in the supplementary materials of this paper, and in Reuning and Plutzer (2020) .

Finally, among the HRS respondents who had non-missing data on this variable and other covariates we employ later, we selected a random sample of 1,000 for our simulations.

The next stage is to simulate actual survey responses to forced choice questions. To do this, we than set all λs equal to 0.8 and drew errors (ε) from a normal distribution with $\sigma_\varepsilon = 0.7$. This produced X_J^* values ranging from about -4.1 to 4.1 which we mapped onto the six answers using cut points of -1, $-.3$, 0, $+.3$, $+1$.

We selected these values to closely mimic the actual responses from the HRS. For example, in the untransformed data, a principal components factor analysis yields a mean λ of 0.74, which is mid-way between our baseline validity and high validity simulations, below. Likewise, the original data have a σ_ε of 0.67, which is similar to our baseline simulation but larger than our high reliability simulation. The simulated data, when combined in a simple additive measure, produces a scale with reliability of 0.85, slightly higher than the actual value of 0.79 because we simplify the data generating process to eliminate stochastic recording errors. Thus, these assumptions of the model are grounded in the example we intend to simulate and the resulting additive scale from the simulated respondents has properties that closely resemble the additive scale from the actual HRS respondents.

Once the dataset was complete, we calculated a series of descriptive statistics including Cronbach's alpha (for the recorded answers X_J) , the total number of straightliners, and the number of straightliners for each possible response (e.g., the number saying "strongly disagree" to all five questions). We then repeated this process 500 times, to have a sampling distribution of outcomes. Simulations were conducted in R (all relevant code is available in the supplemantary materials). The simulation produced what most researchers would consider a good unidimensional scale. Across the 500 simulations the mean value of Cronbach's α is 0.85. If we compute a summary score by adding up the answers, we get a distribution resembling the distribution in the 2006 HRS (Figure 3).

But what about straightlining? In our 500 simulations, the mean percentage of straightliners—giving the identical response to all five questions—was 7.28% (bottom row, Table 1) with 95% of simulations generating straightlining rates between 6.00% and 8.65%. Because of the skew, straightlining was more common for "strongly agree" than other responses.

Recall that the model mimics the data generating process for 1,000 *optimizing* respondents. All generate initial mental answers that correspond well with their latent value ($\lambda = 0.8$); all make some errors, but the variance of the error term is small ($\sigma_\varepsilon^2 = 0.49$) relative to the variance of the latent variable ($\sigma_\eta^2 = 1.00$), and we do not allow any errors at the mapping stage. Therefore, *every straightliner in the simulated data set is a valid straightliner*.

We test two critical aspects of our argument by changing the parameters of the simulation. First, if we repeat the simulation with higher reliability, do we see the expected increase

[1]Wave 6 of the WVS includes grids with these properties for measuring social trust (V102–V107) and social anxiety (V181–V186).

[2]The 2010 health care module of the ISSP includes a version of the Diener scale (IDEALLFE, CONEXCEL, SATLIFE, GOTTHNGS) while the age module includes a worry about aging scale (IMMOBILE, DECIDING, FINDEPND, PAYHLTH, WITHKIDS).

[3]The most recent ANES presidential election study (2016) includes a misogyny scale with these properties (GENDRES_INNOCENT, GENDRES_APPREC, GENDRES_CONTROL, GENDRES_LEASH)

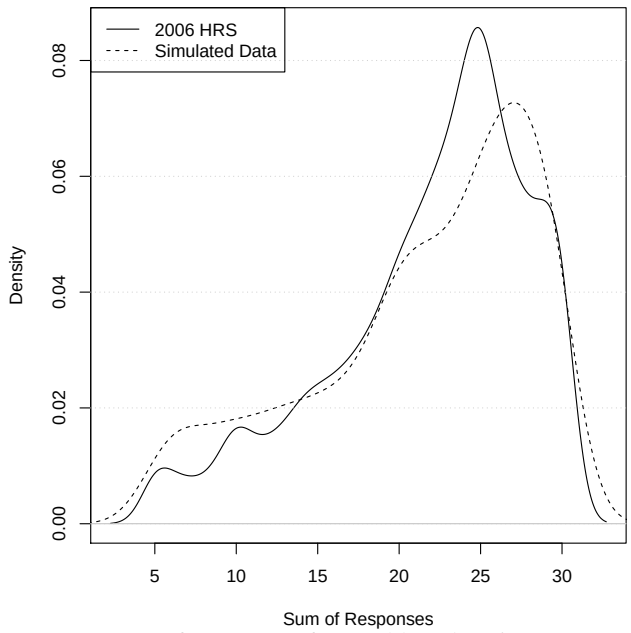

Figure 3. Sum of Responses for Health and Retirement Study (N=7332) and Simulated Data (N=1000)

in straightlining? To do this, we decreased the measurement error's standard deviation from 0.7 to 0.5. The bottom row of Table 1 shows that increasing the reliability *increases* the percentage of straightliners in the average simulated data set from 7.28% to 11.44%.

We next restored σ_ε to 0.7 but increased the validity from 0.8 to 0.9. The results, displayed in the third column of Table 1, show straightlining is again more frequent than in the baseline condition, rising from 7.28% to 10.96%. Interestingly, the distribution of straightlining is different in the cases of Low Validity/High Reliability and High Validity/Low Reliability. When reliability increases, straightlining increases for all six scale points. When validity increases, consistent with the expected wider spread of mental answers, the straightlining mainly increases in the two extreme positions "strongly disagree" and "strongly agree".

Finally, we ran the simulation with both improved validity and reliability. The results, in the last column, reveal a sharp increase in straightlining, now 16.29% of all optimizing respondents, which is almost the same as the percentage as we observe in the actual HRS data set (16.4%). This indicates that a researcher electing to remove straightliners from the dataset might unknowingly remove mainly optimizers.

The results demonstrate that *as data quality increases, so too do the number of straightliners*. High levels of straightlining might actually be a sign that the questionnaire designers did *good* job of writing questions that measure a single unidimensional scale! And in the case of scales with wording consistently in the same direction, the risk of removing valid straightliners will increase as reliability and validity increase.

3.1 Reverse Coding of Responses to Reduce Valid Straightlining

At this point, most readers will be thinking that a grid of questions all in the same direction (as in the Diener scale) stacks the deck towards high levels of valid straightlining. That is indeed the case for responses at the poles of the scale. What is the impact of introducing reverse worded questions? We answer this in Table 2, which reports the baseline results from the original simulations and then reverses the valence of one, and then two, questions.

Column 2 shows that including just one reverse-coded question eliminates 98% of the valid straightlining, going from 73 to less than two respondents in the typical simulated sample of 1,000 optimizers. Making two of the six questions reverse coded further halves the incidence of valid straightlining.

The implication is clear. For Likert-style agree/disagree questions arrayed in a grid, researchers should always include questions worded in opposite directions whenever possible. When this is done, the likelihood of an optimizer selecting "strongly agree" or "strongly disagree" to every question is essentially zero. Indeed, simulations intended to mimic the HRS data, but with one question reverse worded, never produce high levels of straightlining under any combination of validity and reliability.

3.2 Middle Category Straightlining

There is however one special case when reverse wording of questions will not eliminate valid straightlining in a sequence of rating scales. This is when a large percentage of the population have true scores near the midpoint of the latent variable and the response options include a neutral response. This corresponds to the most frequently used agree/disagree format, with five options. A detailed analysis of this possibility is included in the Appendix. This analysis shows that even with reverse worded questions, optimizing respondents can straightline the middle category under common conditions. Table A1 shows that as items become more reliable, middle answer straightlining increases, while increasing validity has the opposite effect. However, in the ideal conditions of high validity and high reliability, the number of valid straightlining the middle category more than doubles compared to baseline levels, increasing from 1.6% to 4.0%. Reverse coding is not necessarily a way to eliminate all valid straightlining but can be a useful tool in situations where there is no middle category and responses are not expected to be centered around it.

Table 1
Percentage of Respondents Straightlining in Simulated Dataset

	Baseline Validity & Baseline Reliability $\lambda = 0.8, \sigma = 0.7$	Baseline Validity & High Reliability $\lambda = 0.8, \sigma = 0.5$	High Validity & Baseline Reliability $\lambda = 0.9, \sigma = 0.7$	High Validity & High Reliability $\lambda = 0.9, \sigma = 0.5$
No Straightlining	92.72	88.56	89.04	83.71
SL Strongly Disagree	1.58	2.35	2.38	3.34
SL Somewhat Disagree	0.11	0.34	0.10	0.34
SL Slightly Disagree	0.00	0.01	0.00	0.01
SL Slightly Agree	0.00	0.02	0.00	0.01
SL Somewhat Agree	0.34	1.22	0.31	1.07
SL Strongly Agree	5.25	7.50	8.17	11.51
Cronbach Alpha	0.85	0.92	0.88	0.94
% Straightlining any answer	7.28	11.44	10.96	16.29

(500 simulations, each with N=1,000)

Table 2
Percentage of Respondents Straightlining, with one or two reverse–worded questions

	Baseline	1 Reverse & Coded Question	2 Reverse & Coded Questions
No Straightlining	92.62	99.86	99.93
SL Strongly Disagree	1.58	0.00	0.00
SL Somewhat Disagree	0.11	0.04	0.02
SL Slightly Disagree	0.00	0.00	0.00
SL Slightly Agree	0.00	0.00	0.00
SL Somewhat Agree	0.34	0.08	0.04
SL Strongly Agree	5.25	0.01	0.00
Cronbach Alpha	0.85	0.85	0.85
% Straightlining any answer	7.28	0.14	0.07

(500 simulations, each with N=1,000)

3.3 The Effects of Removing Valid Straightliners from Analysis

As discussed above one way that researchers use straightlining is to identify respondents to drop from analysis. The logic being that responses from these individuals are unreliable and so removing them will increase the overall data quality. To test the effects of this strategy we introduce a simple, but we think realistic research question. In this, we test the hypothesis that past reports of being subject to discrimination will be associated with lower levels of life satisfaction, after controlling for race, ethnicity, age, education and gender. This is not an especially profound question but is typical of the kinds of questions that social scientists seek to answer with large survey data sets. Our basic strategy is simple. We use our sample of optimizing respondents whose real levels of reported life satisfaction were converted to answers to six simulated questions. Those simulated answers are converted to an additive scale, which is regressed on our independent variable of interest—reports of prior discrimination—and five control variables. The interval variables are all standardized to simplify comparison of their effects.

We estimate this regression model with all straightliners included, and then with them excluded. Figure 4 shows the estimates from the baseline simulations (top panel) and from the high validity and high reliability simulations (bottom panel). We include the other two simulations in the Appendix.

Figure 4 shows the mean estimated coefficients for each of our independent variables and the boxplots show the distribution of the coefficients across the 500 simulations. We begin with the top panel, with baseline levels of validity and reliability ($\lambda = 0.8$, $\sigma_\varepsilon = 0.7$). The boxplots in the top-left corner show that with all straightliners retained in the data set, a one standard deviation change in discrimination decreases life satisfaction by 1.58 (the average across 500

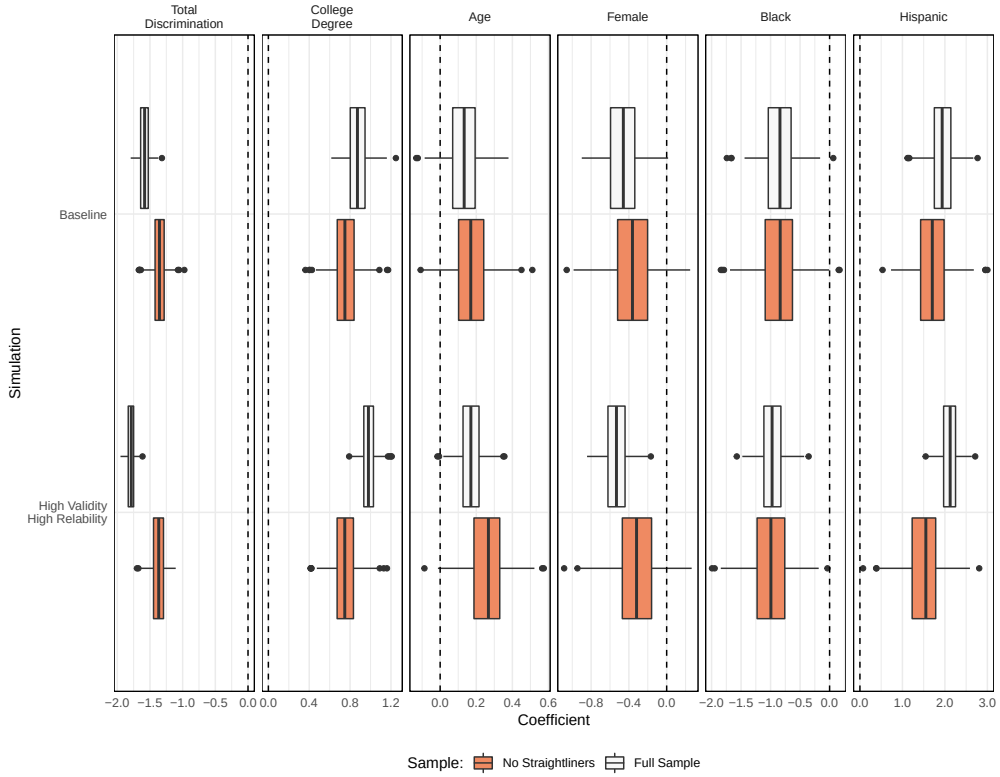

Figure 4. Estimated Coefficients from Simulations

simulations). When the straightliners are excluded, the point estimate is still significant but 13% lower. That is, the removal of 45–95 valid straightliners appears to bias the estimate substantially. For four of the five control variables, the removal or straightlining respondents also shifts the estimated coefficient towards zero, but by very small amounts.

Yet we already know that as reliability and validity improve, the number of straightliners increases. The impact of removing valid straightliners in this situation is assessed in the bottom row. Here we see more dramatic shifts. The effect of being Hispanic on life satisfaction has shrunk by 28% and the effect of gender has shrunk by 39%. The impact of discrimination falls from −1.79 to −1.37—a drop of 24%. Putting these results another way, *whenever researchers remove valid straightliners from a data set, they are inducing a form of sample selection bias*. Researchers increase Total Survey Error by effectively increasing unit non-response (Biemer, 2010).

In answering our primary research question—how much does discrimination impact life satisfaction—this bias is substantial, reducing the apparent effect by 13% to 24%, depending on the overall quality of the data. Shifts like these, even if they do not alter judgments of statistical significance are not trivial. The under-estimated effect of discrimination might, for example, increase the likelihood that scholars and policy analysts perceive daily discrimination and microaggressions as a relatively trivial problem.

4 Illustration #2: Skewed Frequency or Intensity Questions

While the advice to reverse-word some questions is common, many constructs have a "natural" stem direction that makes it difficult to write clear questions in both directions (e.g., requiring the use of confusing double negatives). One important class of such scales are those measuring frequency or intensity relative to a true value of zero. A common example is a battery of questions concerning food insecurity:

We worried whether our food would run out before we got money to buy more. Was that often, sometimes, or never true for you in the last 12 months?

Other examples include self-reports of political participation, incidence of stressful life events, media consumption, and many more. These all have a baseline condition of "never," "none" or similar reference to a true zero.

Of course, some unipolar questions can be easily reversed—for example, "Where you live, how often is the weather sunny?" vs. "Where you live, how often is the weather cloudy?" But many others require double negatives or odd phrasing: it is difficult to formulate a clear question asking individuals how often they *did not* protest. And while it is often fine to present the response categories in either

ascending or descending order, questionnaire designers are reluctant to switch between two presentation orders *within* a set of questions on the same topic.

These unipolar scales can produce high levels of valid straightlining when the distribution of the latent variable is heavily skewed. For example, most people do not regularly participate in political activities while a small portion participate a lot; few people go to bed hungry, but those who do will often experience multiple indicators of food insecurity.

To explore the properties of such unipolar scales measuring highly skewed phenomena, our next simulation is based on the Health and Retirement Survey's everyday discrimination scale (Williams, Yu, Jackson, & Anderson, 1997). The scale's introduction and five questions are: "In your day-to-day life how often have any of the following things happened to you?"

1. You are treated with less courtesy or respect than other people.
2. You receive poorer service than other people at restaurants or stores.
3. People act as if they think you are not smart.
4. People act as if they are afraid of you.
5. You are threatened or harassed.

The HRS version has six answer categories ranging from never to "almost every day." We used the respondents' real answers to generate an approximation of their latent propensity to experience discrimination, and we scaled the latent variable to run from 0 to 31, approximating the number of days each month that each respondent was at risk of experiencing discrimination. This produces a distribution for η that is illustrated in the left-hand panel of Figure 5.

Unreliability: Because the distribution is constrained by zero and thirty-one, residuals cannot be drawn from a symmetric distribution. If the latent variable takes a value of 0, residuals can only be positive (-2 days is not a possible mental answer) and likewise residuals are more likely to be negative when the latent variable takes a high value. To simulate this, we use a Poisson distribution weighted by the distance from the scale's middle point (15) so that errors can be symmetric around the latent value in the middle of the distribution but skewed away from the endpoints. We scale the random error at two levels: 75% and 25% of the simulated Poisson function to reflect low and high reliability conditions.

Validity and Bias: We use a baseline validity coefficient of 0.5, which corresponds to a 50% chance of experiencing any specific kind of discrimination on a day when at risk of discrimination. We apply this baseline differently to create three types of respondents:[4]

Unbiased. Their expected mental answer is equal to $\eta \times \lambda$.

Deniers. These individuals who do not recognize their experiences as discrimination. When they are treated differently than others, they sincerely believe that this is normal

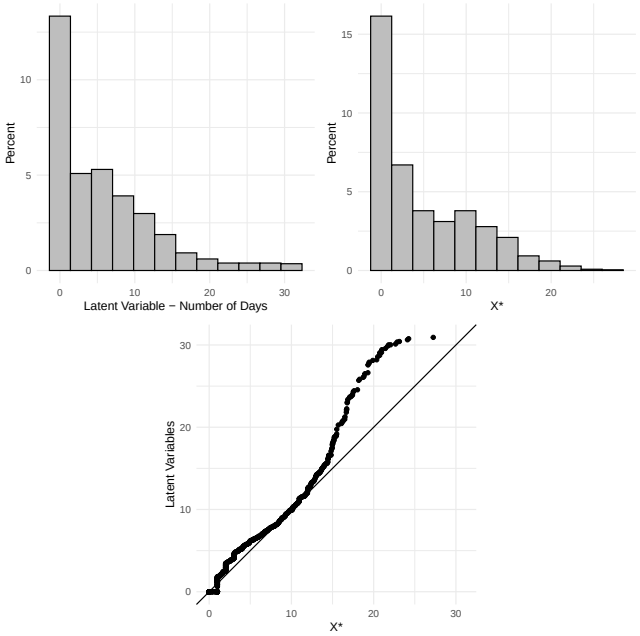

Figure 5. Example Distribution of Everyday Discrimination Latent Variable (η) and Simulated Mental Answer (X^*), Modeled in Low Reliability Condition

everyday activity. They will always report "never." This kind of bias can occur in many such self-reports (Tourangeau, Kreuter, & Eckman, 2012), especially where societal norms many minimize the severity of actions, including self-reports of bullying (D. G. Cornell & Brockenbrough, 2004) and sexual assault (Cantor & Lynch, 2000; Kelly & Stermac, 2008).

Over-reporters. Over-reporting can arise when respondents define many incidents as unfair. For example, they might receive poor service at a badly managed restaurant and feel that they were singled out for poor treatment because of their age, ethnicity, disability, gender or some other trait. They report discrimination even when it does not occur. Over-reporting can also arise if negative experiences are felt more intensely and more likely to be encoded in memory than pleasant or neutral experiences (Strube, 1987). These being more accessible in memory, some optimizing respondents will perceive that negative events happen more frequently than they actually do. We operationalize over-reporting by adding 2 to X^* before translating it into the reported response.

In our simulations, we initially randomly assign 70% of the sample to be unbiased, 15% to be deniers, and 15% over-reporters. The upper right-hand panel of Figure 5 shows the distribution of X^* in the low reliability condition. These show that low reliability pushes questionnaire responses from the

[4]Realistically, physically bullying occurs less often than cyber, verbal and social bullying, so a more realistic distribution would set gamma for the fourth item at about 40% of the value of the others.

Table 3
Two alternative response category options for skewed distribution simulation

OPTION A	
Descriptors	Operationalized as
Never	0
Once or twice	1-2
Three to ten days	3-10
More than ten days	11-31

OPTION B	
Descriptors	Operationalized as
Never	0
Once in a while	1-5
Often	6-20
Every day or almost every day	21-31

highest values to the middle of the distribution.

Cutpoints: We simulate two different verbal signifiers, as illustrated in Table 3. Option B has a much higher level of difficulty to score in the top category.

4.1 Valid Straightlining in Skewed Frequency Simulations

Our initial simulations vary the error variance and the verbal cut points. The upper panel of Table 4 uses response option set "A" with an easy threshold (eleven days or more) for the last category and the lower panel uses option set "B."

More than four in ten optimizers straightline in both reliability conditions across both response sets. With lower reliability, 16.86% say "never" to all four questions. But as data quality improves due to less error variance, 21.85% now straightline the first response category. The effects of improved reliability are complicated by the particularity of the thresholds. For response set A reducing the percentage of individuals who give the highest response to all questions from 11.43% to just 2.61%. For response set B the reduction in straightlining is in the second to last category. Generally, however, the simulations show that improved data quality increases the number of valid straightliners, and not only in the category of "never" where valid straightlining would intuitively be seen as more plausible.

4.2 Eliminating bias

Recall that the simulations reported in Table 4 assume that 15% of the population are deniers and 15% are exaggerators. In studies of bullying, sexual assault, discrimination, and many other topics researchers are concerned with both over- and under-reporting. In the case of bullying, for example, one method thought to reduce over-reporting is to present students with a definition of bullying (*Physical Bullying involves repeatedly hitting, kicking, or shoving someone weaker on purpose*) that emphasizes that conflict and competition among equals does not meet the definition of bullying (Huang & Cornell, 2015).

If such interventions were effective, how would they impact straightlining? To see, we start with the high reliability simulation reported in the lower panel of Table 4, and rerun it with no over-reporters (our population now consists of 85% valid reporters and 15% deniers). Comparing the first and second columns of Table 5 shows that straightlining decreases, but only by 4.3 percentage points. Eliminating exaggerators also changes the distribution of straightlining, with fewer respondents straightlining the higher levels (e.g., straightlining at the Often category went from 18.61% to 16.32% by eliminating overreporting bias). The elimination of over-reporting also increases valid straightlining in the "never" response by 1.5 percentage points.

We then went further and eliminated all deniers as well. The resulting pattern of straightlining appears in the third column of Table 5. The simulations show a substantial decrease in the overall rate of straightlining, from 62.51% to 50.96%. In the absence of systematic deniers, the number of optimizing respondents answering "never" to all four questions drops from over one in five to under one in ten. Conversely there are slight increases in straightlining in the other categories, except for the most extreme which still has no straightlining.

Note also that, with the elimination of systematically biased respondents, reliability has *dropped* slightly, from 0.97 to 0.95. This confirms a well-known feature of many kinds of reporting bias—straightlining, acquiescence bias, primacy and recency effects can all inflate the value of consistency-based measures of reliability (Cronbach's alpha is unbiased only under the assumption that the covariances among all error terms are zero; systematic bias, such as acquiescence bias violates this key assumption; see Lord & Novick, 1968).

4.3 The effect of straightliner removal for frequency scales

The impact of routinely removing those who straightline a series of frequency questions is conditional on two factors. We know of no instances in the literature in which researchers removed respondents consistently answering "never" to a series of similar questions because investigators *intuit* that these are likely valid responses. Removal would be equivalent of converting the model into the second stage of a selection model and the impact on coefficient estimates will depend on whether variables have the same effect on selection (ever experienced discrimination) and on the frequency of experience among those who had. The more likely scenario occurs when a research observes straightlin-

Table 4
Percentage of Respondents Straightlining, with skewed distribution

Response option A		
	Low reliability	High reliability
No Straightlining	56.80	43.08
SL Never	16.86	21.85
SL Once or twice	0.91	4.48
SL Three to ten days	14.00	27.97
SL More than ten days	11.43	2.61
Cronbach Alpha	0.96	0.98
% Straightlining any answer	43.20	56.91

Response option B		
	Low reliability	High reliability
No Straightlining	46.57	37.49
SL Never	16.86	21.85
SL Once in a while	11.51	22.05
SL Often	25.05	18.61
SL Every day or almost every day	0.01	0.00
Cronbach Alpha	0.96	0.97
% Straightlining any answer	53.43	62.51

(500 simulations, each with $N = 1,000$)

Table 5
Percentage of Respondents Straightlining, with skewed distribution and three levels of validity

	High reliability from Table 4B	With no exaggerators	With no exaggerators and no deniers
No Straightlining	37.49	41.69	49.04
SL Never	21.85	23.34	9.85
SL Once in a while	22.05	18.64	21.96
SL Often	18.61	16.32	19.15
SL Almost every day+	0.00	0.00	0.00
Cronbach Alpha	0.97	0.96	0.95
% Straightlining any answer	62.51	58.31	50.96

(500 simulations, each with N=1,000)

ing in other bins and removal's effect depends on the underlying distribution. If the binning leads to most removing valid straightliners with the highest values ("all the time") coefficient estimates can remain unbiased, but the standards errors will inflate not only due to the loss of cases but also due to restricted variation in the independent variable. We report a series of simulations that illustrate these basic principles in Appendix B.

5 Summary

Our simulations mimic the essential features of questionnaire designs used throughout the social, behavioral and health sciences. The effects are generally consistent for Likert style rating questions: First, we find these are highly prone to valid straightlining of all response options when the questions are all worded in the same direction. Second, valid straightlining *increases* with improved data quality. Third, even with reverse worded questions, valid straightlining of the middle response increases as reliability increases, al-

though it decreases with improved validity.

Ordinal scales used to measure frequency of events are often characterized by underlying skewed distributions, leading many optimizing respondents to straightline the lowest (e.g., "never") category. Further, the number of valid straightliners in other categories is highly dependent on the cutpoints implied by verbal signifiers (e.g., "rarely," "often"). In our simulations we found that improved reliability (at the level of the mental answer) increased the incidence of valid straightlining—much like we found in the simulations of rating scales.

When we simulated the elimination of over- and under-reporting, we saw that the impact on the number of valid straightliners varied: eliminating under-reporting naturally reduces the number of respondents saying "never" to all four questions; but in shifting to other answers, there will still be some increase in the number of valid straightliners.

6 Discussion

Kim et al. (2019, p. 2) argued "Understanding straightlining behavior is important because it may deteriorate both reliability and validity of survey responses." This is only true when it is the result of satisficing, but in many common situations straightlining is a consequence of increasing data quality. Critically, straightlining is not an *indicator* of reliability or validity. As we have shown, reliability and validity have causal effects on the frequency of valid straightlining, and frequently in a counter-intuitive direction.

Our findings have several important implications for practice and for future research priorities. At the design stage, investigators should design question sets in ways that minimize the prevalence of valid straightlining. This includes (a) reverse wording at least one question in a set of attitude questions whenever possible, (b) devising ways to reduce the number of valid answers in single bins—for example by subdividing response categories, or adding follow-up questions in a branching pattern, (c) when asking about behaviors on an ordinal scale, consider branching questions or intersperse questions with different number of response categories. For example, an initial yes/no screening question about having ever experienced a form of discrimination could branch those answering "yes" to a series of questions relevant to them. When feasible, (d) include attention checks or trap questions to provide independent measures of satisficing. These recommendations comport with existing best practices in questionnaire design, and our simulations simply provide additional motivation for adopting them.

At the analysis stage (for already collected data) we recommend against the systematic removal of apparent straightliners. Doing so is akin to classifying them as non-respondents for the purpose of a particular analysis, and if a large number are removed, design and post-stratification weights may be rendered invalid.

Instead, data analysts can use the methods we have introduced here to see if valid straightlining is expected to be prevalent, and to see if their question sets contain features that systematically produce valid straightlining (e.g., low validity leading to heaping in the middle category). If valid straightlining seems unlikely, we recommend that investigators utilize other available indicators of satisficing, such as speeding, to flag respondents who are most likely to be invalid straightliners.

If investigators are tempted to drop straightliners from their analysis data set, we recommend that researches undertake sensitivity analyses along the lines of the regression estimates we presented earlier. If many key estimates shift towards zero after straightliners are removed, that is a clue that the discarded group contains a large number of optimizing respondents whose answers happen to coincide.

Finally, we recommend against case removal as the standard solution. One simple alternative is to recode straightliners' answers in a question set to missing values and use multiple imputation methods to retain the respondents and all the additional valid answers they may have provided. The assumption that the *specific questions* are Missing at Random will always be at least as plausible as the assumption that *the entire unit* is MAR. So imputing specific answers requires weaker assumptions than deleting entire cases. Imputation has the added benefit that researchers need not recalculate analysis weights.

Unfortunately, in many instances—such as the Health and Retirement Survey—there will be reason to believe that the group of straightliners is a mixture of valid straightliners and satisficers. This kind of situation highlights the value of conducting simulations of the kind we described here. Researchers can adapt our simulation approach to capture features of their own data sets (e.g., the number of questions, number of reverse worded questions, number of response categories) and to approximate a range of plausible (but unknown) features—such as the underlying validity, reliability, and the kinds of biased responding indicated by the substantive and methodological literature. Researchers can then approach data analysis with some best- and worst-case estimates of the expected number of straightliners who are actually optimizing respondents.

Looking ahead, survey scientists should strive to develop methods that would allow more confident identification of satisficers and optimizers. For example, latent class mixture models may be a promising method to estimate the probability that any individual straightliner is a valid responder or a satisficer. Another promising approach could involve the application of multiple over-imputation (Blackwell, Honaker, & King, 2017a, 2017b). Multiple over-imputation uses Bayesian shrinkage to create plausible values lying between the recorded value (suspected of the result of invalid straightlining) and plausible values calculated on

the assumption that the data are missing at random. This seems well suited to the analysis of datasets containing a mixture of valid and satisficing straightliners. These agendas may be challenging, but every journey begins with small steps. We believe that the conceptual formalization of valid straightlining and the demonstration of common circumstances giving rise to valid straightlining is one such step.

References

Bethlehem, J., & Biffignandi, S. (2012). *Handbook of web surveys*. Hoboken, NJ: John Wiley & Sons.

Biemer, P. P. (2010). Overview of design issues: Total survey error. *Handbook of survey research*, *2*, 27–57.

Blackwell, M., Honaker, J., & King, G. (2017a). A unified approach to measurement error and missing data: Details and extensions. *Sociological Methods & Research*, *46*(3), 342–369.

Blackwell, M., Honaker, J., & King, G. (2017b). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research*, *46*(3), 303–341.

Bohrnstedt, G. W. (2010). Measurement models for survey research. *Handbook of survey research*, *2*, 347–404.

Cantor, D., & Lynch, J. P. (2000). Self-report surveys as measures of crime and criminal victimization. *Criminal justice*, *4*(2000), 85–138.

Cole, J. S., McCormick, A. C., & Gonyea, R. M. (2012). *Respondent use of straight-lining as a response strategy in education survey research: Prevalence and implications*. Paper presented at American Educational Research Association Annual Meeting.

Cornell, D., Klein, J., Konold, T., & Huang, F. (2012). Effects of validity screening items on adolescent survey data. *Psychological assessment*, *24*(1), 21.

Cornell, D. G., & Brockenbrough, K. (2004). Identification of bullies and victims: A comparison of methods. *Journal of School Violence*, *3*(2-3), 63–87.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, *49*(1), 71–75.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370–392.

Greszki, R., Meyer, M., & Schoen, H. (2014). The impact of speeding on data quality in nonprobability and freshly recruited probability-based online panels. *Online panel research: A data quality perspective*, 238–262.

Health and Retirement Survey. (2014). Data from wave 6, release version from december 2014. Retrieved from https://hrs.isr.umich.edu/

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

Huang, F. L., & Cornell, D. G. (2015). The impact of definition and question order on the prevalence of bullying victimization using student self-reports. *Psychological assessment*, *27*(4), 1484.

Kelly, T. C., & Stermac, L. (2008). Underreporting in sexual assault: A review of explanatory factors. *Baltic Journal of Psychology*, *9*(1/2), 30–45.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233.

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

Lieberman, B., Hancuch, K., & Buttermore, N. (2019). If you're extremely satisfied are you completely satisfied? measuring the relative distance between verbal labels on a response scale., 74–94. doi:10.4324/9781003061601-6

Lord, F., & Novick, M. (1968). Statistical theories of mental test scores. edited by. Reading, MA. Addison-Wesley Publishing Company.

Reuning, K., & Plutzer, E. (2020). *Replication materials for 'valid vs. invalid straightlining: The complex relationship between straightlining and data quality'*. Harvard Dataverse, V1, UNF:6:k+GcDDy5lVzd88V7+ZFDyA==. doi:10.7910/DVN/WE6T71

Saris, W., Revilla, M. A., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with construct-specific response options. *Survey Research Methods. 2010; 4 (1): 61-79. DOI: 10.18148/srm/2010. v4i1. 2682*.

Schonlau, M., & Toepoel, V. (2015). Straightlining in web survey panels over time. In *Survey research methods* (Vol. 9, pp. 125–137).

Strube, G. (1987). Answering survey questions: The role of memory. In *Social information processing and survey methodology* (pp. 86–101). Springer.

Tourangeau, R., Kreuter, F., & Eckman, S. (2012). Motivated underreporting in screening interviews. *Public Opinion Quarterly*, *76*(3), 453–469.

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

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

Williams, D. R., Yu, Y., Jackson, J. S., & Anderson, N. B. (1997). Racial differences in physical and mental health: Socio-economic status, stress and discrimination. *Journal of health psychology*, *2*(3), 335–351.

Yan, T. (2008). Nondifferentiation. *Encyclopedia of Survey Research Methods*, *2*, 520–521.

Zaller, J. R. et al. (1992). *The nature and origins of mass opinion*. Cambridge university press.

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. In *Survey research methods* (Vol. 8, pp. 127–135).

Appendix A
Simulation of Middle Category Straightlining

There is one special case when reverse wording of questions will not eliminate valid straightlining in a sequence of rating scales. This is when a large percentage of the population have true scores near the midpoint of the latent variable and the response options include a neutral response. This corresponds to the most frequently used agree/disagree format, with five options.

To explore this, we return to our simulated diener scale but subtract 0.6 from it so that it is centered around 0. We keep our baseline validity and reliability measures and set cutpoints at -1, -0.5, +0.5 and +1. These cutpoints push more of the responses towards the middle category so we can better examine how straightlining the middle category varies. We then manipulate validity for low (λ = 0.5), baseline (λ = 0.8) and high (λ = 0.9); and reliability as before (σ_ε = 0.5 and 0.7). The results are in Table A1. The first four columns repeat the same comparisons examined in Tables 1 and 2, while columns five and six add a new, low validity, condition.

In the baseline condition (first column), the total rate of straightlining is similar to what we find with six answer options. We also see that increasing the reliability of the data (column 2) again increases the number of valid straightliners in every single response option. However, the biggest increases are found in the neutral category (typically "neither agree nor disagree"). This increase will not be eliminated by reverse wording one or more questions.

The last two columns of Table A2 show that low validity concentrates mental answers toward the middle of the distribution, and therefore contributes to straightlining of the neutral category.

As a result, we can conclude that, in a sample consisting entirely of optimizers, patterns of straightlining of the middle category can result from both poor data quality (low validity) and high data quality (high reliability). In situations like this, removing straightlining respondents would, again, result in the removal of optimizers who did their best to provide thoughtful answers.

To confirm that reverse coding does not eliminate straightlining of the middle category, we re-ran the simulation first with one, and then with two, reverse-worded questions, as reported in Table A2. The first column reproduces column four from the previous table (the high validity and high reliability condition). As before, reverse-worded questions completely eliminate straightlining the endpoints, but does nothing to reduce straightlining the neutral response.

The implications for data analysis are important. In a data set of 1,000 respondents, we might expect 40 valid straightliners of the middle category. Analysts who suspect that the offering of a neutral response might spur status-quo satisficing might be tempting to remove these respondents. But that would entail throwing out perfectly good data produced by optimizing respondents.

Table A1
Percentage of Respondents Straightlining with neutral option

	Baseline Validity & Baseline Reliability $\lambda = 0.8, \sigma = 0.7$	Baseline Validity & High Reliability $\lambda = 0.8, \sigma = 0.5$	High Validity & Baseline Reliability $\lambda = 0.9, \sigma = 0.7$	High Validity & High Reliability $\lambda = 0.9, \sigma = 0.5$	Low Validity & Baseline Reliability $\lambda = 0.5, \sigma = 0.7$	Low Validity & High Reliability $\lambda = 0.5, \sigma = 0.5$
No Straightlining	93.25	88.17	91.13	85.79	96.99	91.98
SL Response 1	3.72	5.03	5.03	6.53	0.53	0.76
SL Response 2	0.03	0.10	0.03	0.09	0.03	0.10
SL Response 3	1.56	4.52	1.40	3.99	2.28	6.91
SL Response 4	0.06	0.22	0.06	0.23	0.05	0.17
SL Response 5	1.40	1.96	2.35	3.36	0.12	0.08
Cronbach Alpha	0.84	0.90	0.86	0.92	0.69	0.80
% Straightlining any answer	6.75	11.83	8.87	14.21	3.01	8.01

(500 simulations, each with N=1,000)

Table A2
Percentage of Respondents Straightlining with neutral option, with and without reverse worded questions

	High Validity & High Reliability $\lambda = 0.9, \sigma = 0.5$	1 Reverse Coded Question $\lambda = 0.9, \sigma = 0.5$	2 Reverse Coded Question $\lambda = 0.9, \sigma = 0.5$
No Straightlining	85.79	95.98	95.99
SL Response 1	6.53	0.00	0.00
SL Response 2	0.09	0.01	0.00
SL Response 3	3.99	4.00	4.01
SL Response 4	0.23	0.01	0.00
SL Response 5	3.36	0.00	0.00
Cronbach Alpha	0.92	0.92	0.92
% Straightlining any answer	14.21	4.01	4.01

(500 simulations, each with N=1,000)

Appendix B
Simulation of a Skewed Distribution

To estimate a skewed distribution we use the everyday discrimination scale. This was adapted by HRS from Williams et al. (1997). The question wording is:

In your day-to-day life how often have any of the following things happened to you?

30a You are treated with less courtesy or respect than other people

1 Almost every day
2 At least once a week
3 A few times a month
4 A few times a year
5 Less than once a year
6 Never

30b You receive poorer service than other people at restaurants or stores

1 Almost every day
2 At least once a week
3 A few times a month
4 A few times a year
5 Less than once a year
6 Never

30c People act as if they think you are not smart

1 Almost every day
2 At least once a week
3 A few times a month
4 A few times a year
5 Less than once a year
6 Never

30d People act as if they are afraid of you.

1 Almost every day
2 At least once a week
3 A few times a month
4 A few times a year
5 Less than once a year
6 Never

30e You are threatened or harassed.

1 Almost every day
2 At least once a week
3 A few times a month
4 A few times a year
5 Less than once a year
6 Never

In order to examine how the removal of straightliners impacts regression analysis we estimated a linear regression model predicting life satisfaction using the simulated discrimination scales as the main independent variable. We also included indicators for if a respondent had a college degree, and if they were female, black or Hispanic along with the respondents age. For each simulation we estimated three models, one with all the data, one with all straightliners removed no matter what they straight lined and one where we removed straightliners that straightlined something beyond the "Never" response (we refer to this last sample as "no positive straightliners").

Figure B1 shows the estimated distribution of the coefficients on discrimination for the three different samples across the range of simulation setups. The important comparison is how the orange (no straightliners) and the green (no positive straightliners) boxplots vary from the blue (full sample) boxplot.

Under both removal rules, removing straightliners the variation in the coefficients increases. Removing all the straightliners tends to inflate the estimates of the coefficients. This is a result of removing a large proportion of individuals who straight lined the "never" category. In contrast when only those who straight lined the other categories is removed the coefficients tend to shrink towards zero. The change is not as extreme as in other simulations, but still poses a threat to inference.

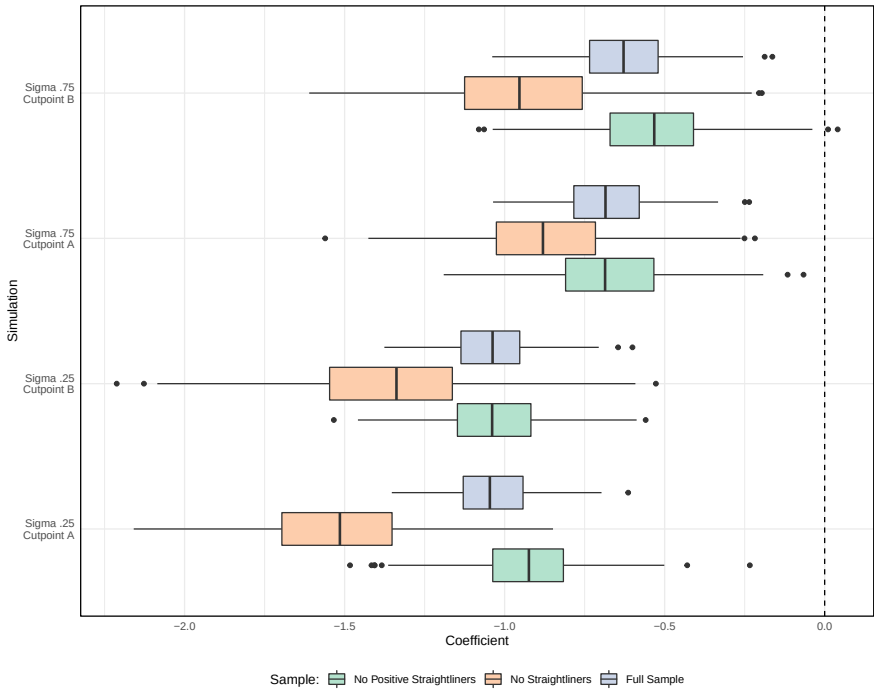

Figure B1. Boxplot of Coefficients on Discrimination from Simulations

Appendix C
Stata Code for Generation of η_i for Simulations

Diener Life Satisfaction Scale

The latent variable is generated from respondents' actual answers to the six original items, which have then been smoothed and jittered to more closely resemble a continuous latent variable.

Step 1 Generate standard summative scale

```
. gen diener_total = klb003a + klb003b +
klb003c + klb003d + klb003e
```

Step 2 Adjust for heaping at scores of 10 and 25

```
. gen diener_smooth = diener_total
. replace diener_smooth = diener_smooth
+ (.25*zscore) if diener_smooth==10
. replace diener_smooth = diener_smooth
+ (.25*zscore) if diener_smooth==25
```

Step 3 Then spread out scores of 5 (the minimum score possible) to eliminate floor effects

```
. replace diener_smooth = diener_smooth
+ weib_left if diener_smooth==5
```

Step 4 Then spread out scores of 30 (the maximum score possible) to eliminate ceiling effects. This is imputed in proportion to respondents wellness level as indicated by their score on the depression scale.

Step 4a First calculate an adjustment component

```
. gen depress_adjust = 0
. replace depress_adjust = (15 -
depress_total)/3 if depress_total==15
```

Step 4b Then add adjustment to scale score for randomly selected 75% with maximum score on the depression scale.

```
. replace diener_smooth = diener_smooth
+ depress_adjust if diener_total == 30 &
uni_0_1 > .25
```

Step 4c Then add adjustment to scale score for randomly selected 25% with one below maximum score on the Mroczek & Kolarz positive/negative affect scale.

```
. replace diener_smooth = diener_smooth
+ depress_adjust if diener_total == 29 &
uni_0_1 == .25
```

Step 5 Then slight jitter to smooth the distribution.

```
. replace diener_smooth = diener_smooth
+ (.2*zscore)
```

Everyday discrimination scale

Step 1 Reverse code from 6 to 1, to 0 to 5.

```
. foreach var of varlist klb030*{
      . replace var' = (var'*-1)+6
. }
```

Step 2 Standardize each item.

```
. egen a_sd = sd(klb030a)
. egen b_sd = sd(klb030b)
. egen c_sd = sd(klb030c)
. egen d_sd = sd(klb030d)
. egen e_sd = sd(klb030e)
. gen disc_a_z = klb030a/a_sd
. gen disc_b_z = klb030b/b_sd
. gen disc_c_z = klb030c/c_sd
. gen disc_d_z = klb030d/d_sd
. gen disc_e_z = klb030e/e_sd
```

Step 3 Create summative scale of standardized items.

```
. egen discrim_eta0 = rowmean(disc_*_z)
```

Step 4 Trim values in tail to that of 99%-tile.

```
. replace discrim_eta0 = 3 if
discrim_eta0 > 3 & discrim_eta0==.
```

Step 5 Calculate jitter

```
. gen uni_0_fifth = runiform(0,.2) - .1
. replace uni_0_fifth = 0 if
discrim_eta0==0
```

Step 6 Add jitter and multiply by ten to convert to number of days per month.

```
. gen discrim_eta = (discrim_eta0 +
uni_0_fifth)*10
```