Examining Noneffortful Responding in Survey Responses and SEL Measures

Rik Lamm, Mohammed A. A. Abulela, Kyle Nickodem, Michael C. Rodriguez
University of Minnesota

Minnesota Youth Development Research Group
www.mnydrg.com

April 14, 2024

Paper presented at the annual meeting of the
National Council on Measurement in Education, Philadelphia, PA.

Citation:

Lamm, R., Abulela, M. A. A., Nickodem, K., & Rodriguez, M. C. (2024, April 14). *Examining noneffortful responding in survey responses and SEL measures* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Philadelphia, PA. https://hdl.handle.net/11299/194886

Examining Noneffortful Responding in Survey Responses and SEL Measures

Background

Self-report measures or surveys have been one of the most commonly used methods to gather information in educational and psychological research (Arias et al., 2020). Surveys have been used increasingly in K-12 settings (Soland et al., 2019). However, given that most surveys are administered in low-stakes contexts, noneffortful responding (NER; selecting a response option without paying attention to the item content) has been identified as a threat to the validity of score-based inferences (van Laar & Braeken, 2022). Thus, survey developers and measurement specialists are paying more attention to this test-taking behavior in an attempt to reduce its negative impact on the quality of data and on data-informed decisions (Goldammer et al., 2020).

In this paper, we briefly summarized what is known about NER prevalence, its impact on psychometric properties of measurement instruments, and detection methods. To examine these topics in the context of noncognitive measures, specifically regarding three measures of social and emotional competencies, we analyzed data from a large-scale statewide youth development survey, identified NER, and evaluated the impact of removing NER cases.

Prevalence of NER

NER has been prevalent in low-stakes large-scale surveys administered in national and international assessments such as TIMSS and PISA. In cross-sectional and experimental research, various researchers have found varying amounts of NER in their data.

For instance, Oppenheimer et al. (2009) conducted two experimental studies and indetified NER using an instructional manipulation check (IMC), which assessed whether respondents read instructions as an indirect measure of satisficing. In the first study, 213 college students (156 women, 57 men) were involved, of which 46% failed the IMC and consequently were identified as careless responders. Specifically, participants were presented with a scenario about attending a game and their response indicated whether they read the instructions carefully. In the second study, 144 (76 women, 68 men) college students participated, of which 50 or 35% failed the IMC. In addition to the scenario manipulated in the first study, participants completed a 7-item abbreviated version of the maximization scale, which assessed respondents' tendancy to maximize versus satisfice (also see Böckenholt, 2017; Calderon Vriesema & Gehlbach, 2021).

Maniaci and Rogge (2014) collected data from 674 adults (over 18 years old) who responded to a 44-item online survey assessing the Big Five Personality traits in addition to the Attentive Responding Scale (33 items) and the Directed Questions Scale (7 items). To detect NER, the authors utilized the long string index, multivariate outliers, even–odd consistency, and psychometric synonyms and antonyms. Overall, the authors found that NER ranged from 3% to 9% of respondents.

On a social and emotional learning measure, Steedle et al. (2019) utilized a sample of 18,578 high school students who took ACT Engage (ACT, 2016) between 2009 and 2013. The authors utilized nine NER detection methods and found that individual methods indicated that NER ranged approximately from 1% to 20%.

To summarize, the rate of NER prevalence has varied as a function of the context as well as how researchers operationalized NER.

Impact of NER on Survey Psychometrics

NER has been found to negatively impact psychometric properties of measurement instruments and consequently threaten the validity of score-based inferences. Briefly, NER has been found to distort factor analytic results, including item loadings and associated fit indices (e.g., Woods, 2006). NER has been found to distort validity evidence based on associations with other variables (e.g., Schneider et al., 2018) and underestimate score reliability estimates (Arias et al., 2020). To summarize, NER impacts the quality of validity evidence as well as score reliability coefficients.

Detection of NER

Since a small rate of NER has been found to impact item and survey psychometric properties, several methods have been proposed to detect this undesired test-taking behavior. The methods utilized to detect NER include inter-item standard deviation (IDS), long string index, and odd-even consistency index. The IDS is a within-person standard deviation, which refers to the standard deviation computed for all items for each respondent (Marjanovic et al., 2015). The long string index has been utilized when an individual consecutively selects the same response category across many items (Goldammer et al., 2020). The odd-even consistency index is computed by dividing the items of a scale into odd and even subsets, followed by computing the score for each subset and then estimating the correlation coefficient between the two subsets (Huang et al., 2012). Since the odd-even consistency index is a function of correlating pairs of item responses within person, and the measures used in this study are brief, this index would be based on too few data points to be meaningful and was not employed. For a comprehensive presentation of detection methods, interested readers are referred to Curran (2016).

Purpose of Curent Study

To date, the impact of NER on social and emotional learning (SEL) competencies has not been investigated in a statewide survey that includes responses provided by students with different characteristics. To understand the potential impact of NER in such a case, we pursued three research questions:

1. What is the proportion of respondents engaging in NER using two criteria for detection: inter-item standard deviation measuring individual consistency and long string measuring invariability of responses?
2. Does the proportion of NER vary across grade levels?
3. Does removal of noneffortful respondents alter confirmatory factor analysis model-data fit, item parameter estimates, or score descriptive statistics?

Methods

Data Source and Measures

The data came from the Minnesota Student Survey, a statewide survey asking students about non-academic topics associated with their experiences in schools and communities (Minnesota Department of Education, 2024). The data used in the study came from the 2022 administration of the survey, with a total sample size of 135,447 with approximately proportional sample sizes for grades 5, 8, 9, and 11.

From the survey, multiple measures of developmental skills and supports were estimated that were the focal point for this study. These measures were based on the developmental asset framework introduced by Search Institute (2024; Scales & Leffert, 2004). The first two measures were considered developmental skills or intrapersonal attribute. The third measure was considered a developmental support or interpersonal attribute. All three measures were originally developed for the Developmental Asset Profile (Search Institute, 2013) and were based on a deep research foundation grounded in positive youth development (Benson et al., 2006).

Positive Identity and Outlook (PIO) is a six item measure about feeling good about one's future and positively handling difficult situations (e.g., “I feel good about myself” and “I find good ways to deal with things that are hard in my life”).

Social Competence (SC) is an 8-item measure about decision-making skills for building relationships (e.g., “I build friendships with other people” and “I resolve conflicts without anyone getting hurt”). The Positive Identity and Outlook and Social Competence measures use a 4-point response scale ranging from *Not at all* or *Rarely* to *Extremely* or *Almost always.*

Empowerment (EM) is a 6-item measure regarding feeling valued and safe. Three items (e.g., “I feel safe at home”) used a 4-point *Strongly agree* to *Strongly disagree* scale, and three items (e.g., “I feel valued and appreciated by others”), used a 4-point response scale ranging from *Not at all* or *Rarely* to *Extremely* or *Almost always*.

These three measures were imbedded in a much longer survey, in total including over 100 items in the elementary school version and over 200 items in the high school version (although many of these other items were not administered, for example, depending on the extent to which students engaged in risky behavior, such as substance use). Each measure was Rasch scaled for analysis purposes and were found to have adequate fit in confirmatory factor analysis to support the use of the Rasch model; the measures were particularly of high quality in terms of little to no differential item functioning regarding race, gender, and grade (Rodriguez, 2023).

Criterion for NER Identification

Inter-item standard deviation (IDS) was estimated by calculating the standard deviations for each individual’s response patterns, for the items of interest. The items for each of the skills and supports were measured from the respective scale scores. For each individual, a standard deviation and associated variance were calculated. An IDS above 1.0 was used as an indicator of NER. Individuals who achieved this level or higher were identified as engaging in NER.

Long string index was estimated by counting the number of identical responses in a row, for each individual. From the survey, there were 20 items that comprised the three skills and supports of interest and thus, selecting over 10 responses in a row identically indicated NER.

Removal of NER

For each of the three factors, confirmatory factor analysis was performed to check the quality of the factor. Following this, CFA was conducted for each factor after the removal of non-effortful responses from each method. The factors were compared between the original factor and the modified factors using the fit statistics of root mean squared error of approximation (RMSEA), Comparative Fit Index (CFI), and Baysian Information Criterion (BIC).

RMSEA is a measure of the amount of error in the fit, and a proper fit should be below .08 for reasonable model fit (Hu & Benter, 1999). CFI is a measure of fit of the target model to an independent or null model. A value above .90 is associated with proper fit. BIC is a relative fit index which is derived from the log likelihood of the model, with a constraint based on sample size. A lower BIC is associated with better fit for comparable models.

Results

With the first method of NER identification (inter-item variances > 1.0), 10% (13,801) of students had a variance greater than 1.0 for the target items. These students had highly varying response patterns, such as choosing *strongly disagree* and *strongly agree* for items that have been shown to be measuring the same factor. This potentially indicated that the student was not engaging with the survey and likely choosing response options randomly.

The next method for examining NER, long string analysis, showed that 6% (8,012) of the respondents showed a string of identical responses for over half of the items. This is an indicator of NER; however, it is less of a strong indicator due to factors including items that are similar, measuring the same construct, and thus, should be answered similarly. Although, long string analysis does indicate that a student may not be fully engaging with each of the items and answering the items quickly, without considering each one in a unique manner.

The evidence of NER differed by grade for each of the methods. The inter-item variances method resulted in slight differences between the grades, with grade 9 having the lowest amount of NER at 10.8% (Table 1). Grade 5 had the highest level of this method of NER at 13.5%. There was not an overall consistent trend across grades.

The next method, long string analysis, resulted in more variation between the grades with the lowest percent being for grade 5 at 3% and the largest for grade 8 at 7.3% (over double). Overall, the higher grades have substantially higher percentages for NER in this method than for grade 5. This is evidence that students in elementary schools are more careful with their responses and are less likely to answer every item identically.

Table 1

Number and Percentages for Non-Effortful Responses by Detection Method and Grade

	Grade			
Method	5	8	9	11
Inter-item variances > 1.0				
n	7915	7500	6770	5180
%	13.5%	11.1%	10.8%	11.5%
Long string > half				
n	1030	2772	2347	1863
%	3.0%	7.3%	6.6%	6.8%

Note. Total $N = 135{,}447$. Due to missingness, totals may not be equal.

The fit statistics for each of the models were presented in Table 2. Overall, each factor had modest to adequate fit. For Positive Identity and Outlook, the fit statistics were different based on the NER detection method. Using RMSEA, the removal of long string responses resulted in improved fit the most. For Social Competency, removing the responses with large item variances provided the best fit for RMSEA and CFI. Empowerment results were similar to PIO, with long string being most affected regarding RMSEA. Overall, the fit for each factor was improved from the removal of NERs.

Table 2

Fit Statistics for CFA Models, Given Removal of NERs by Detection Method

NER detection method	Fit measure		
	RMSEA	CFI	BIC
Positive Identity and Outlook			
Full model	.133	.923	1179557
Item variances	.131	.927	1012811
Long string	.130	.914	1094595
Social Competence			
Full model	.100	.891	2110449
Item variances	.093	.916	1798664
Long string	.097	.878	1983609
Empowerment			
Full model	.185	.803	1316677
Item variances	.184	.824	1107364
Long string	.180	.796	1236759

Table 3 contains the standardized item loadings for each factor, for each of the detection methods. This permits identification of differences in item parameter estimates between the NER methods, allowing us to examine whether the items changed in their parameter estimates greatly. For PIO, the item loadings were similar across methods. The exceptions were for items 60g and 60n, with the long string method resulting in much lower estimates. Important to note is that none of the estimates changed enough to re-order the items. Overall, the larger factor loadings resulted from the item variances method and the lowest were seen in the long string method.

Similar to PIO, the SC factor has similar loadings across NER removal methods. Additionally, the order of the items did not change, except for a small decrease in 60k. Again, the item variances method showed the largest factor loadings. The EM factor items also did not change in order and had larger factor loadings for the item variances NER.

Table 3

Standardized item loadings for CFA Models, Given Removal of NERs by Detection Method

Items within factor			NER detection method	
		Full model	Item variances	Long string
Positive Identity and Outlook				
60a	I feel in control of my life and future.	.76	.75	.75
60b	I feel good about myself.	.76	.75	.74
60f	I feel good about my future.	.79	.79	.78
60g	I deal with disappointment without getting too upset.	.52	.55	.44
60h	I find good ways to deal with things that are hard in my life.	.72	.73	.67
60n	I am thinking about what my purpose is in life.	.18	.25	.11
Social Competence				
60c	I say no to things that are dangerous or unhealthy.	.62	.63	.61
60d	I build friendships with other people.	.47	.50	.41
60e	I express my feelings in proper ways.	.60	.63	.56
60i	I plan ahead and make good choices.	.67	.69	.63
60j	I stay away from bad influences.	.69	.70	.68
60k	I resolve conflicts without anyone getting hurt.	.63	.66	.60
60m	I accept people who are different from me.	.37	.41	.34
60q	I am sensitive to the needs and feelings of others.	.43	.47	.39
Empowerment				
22b	I feel safe at school	.47	.50	.46
22c	I feel safe in my neighborhood	.49	.54	.49
22d	I feel safe at home	.52	.55	.53
60l	I feel valued and appreciated by others	.66	.68	.63
60o	I am included in family tasks and decisions	.72	.72	.69
60p	I am given useful roles and responsibilities	.71	.72	.68

The final component to examine the effect of NER was the review of the summary statistics from the scores obtained for each factor. After Rasch scaling, scores were obtained for each individual on each factor. Scores were centered at 10, with that level being neither "like me" nor "dislike me" (to support interpretation)—in this way, scores larger than 10 were viewed as positive. The means and standard deviations for each factor were provided in Table 4 and compared across NER method. Similar to the factor loadings, the means and standard deviations were similar within each factor, when comparing NER method. For each, the highest mean scores were generated from the item variances method and the lowest scores were from the long string method. The largest change in means, from the full model as a referent, was seen in Empowerment, with a standardized mean difference (*d*) of 0.29 following the remove of NER with the item variances method, potentially a nonignorable difference.

Table 4

Summary Statistics for Rasch Scores, Given Removal of NERs by Detection Method

NER detection method	***M***	***SD***	***d***
Positive Identity and Outlook			
Full model	10.29	1.02	
Item variances	10.43	1.14	-0.14
Long string	10.20	0.87	0.09
Social Competence			
Full model	10.67	0.89	
Item variances	10.84	1.02	-0.19
Long string	10.58	0.77	0.10
Empowerment			
Full model	11.42	1.18	
Item variances	11.76	1.36	-0.29
Long string	11.32	1.10	0.09

Note. Cohen's *d* was estimated comparing *M*s with the full model, with the full model *SD* as the denominator.

Conclusion and Implications

Overall, the level of NER for social and emotional learning measures in the MSS was within the range seen among other surveys and assessments of this length. Overall, the MSS appears to be a quality tool that keeps survey respondents engaged in the items. Additionally, the results from the survey were modestly impacted by NER.

We observed some differences in NER between grades and across three measures of SEL. Important to note is that the fit for all measures improved after the removal of NERs, although only marginally. Additionally, neither of the NER methods was consistently better across contexts. Similarly, for the item parameters and score parameters, there was little difference between the methods.

We acknowledge that addressing NER must be done with caution. Although the indicators of NER are quantitative, the criteria at which each indicator might support a decision to remove a response or a participant as noneffortful requires human judgement. We encourage additional research to evaluate the impact of different levels of criteria to make such decisions.

References

ACT. (2016). *ACT Engage user guide*. https://www.act.org/content/dam/act/unsecured/documents/EngageUserGuide.pdf

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods, 52*(6), 2489-2505. https://doi.org/10.3758/s13428-020-01401-8

Benson, P. L. (1990). *The troubled journey: A portrait of 6th to 12th grade youth*. Search Institute. http://pub.search-institute.org/file/archive/1990-Benson-Troubled-Journey.pdf

Benson, P. L., Scales, P. C., Hamilton, S. F., & Sesma, A. (2006). Positive youth development: Theory, research, and applications. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol. 1* (6th ed., pp. 894-941). John Wiley & Sons.

Calderon Vriesema, C., & Gehlbach, H. (2021). Assessing survey satisficing: The impact of unmotivated questionnaire responding on data quality. *Educational Researcher, 50*(9), 618-627. https://doi.org/10.3102/0013189X211040054

Curran, P. G. (2016). Methods for detection of carelessly invalid responses in survey data. *Journal of Experimental Psychology, 66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly, 31*(4), 1-16. https://doi.org/10.1016/j.leaqua.2020.101384

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*, 99–114. https://doi.org/10.1007/s10869-011-9231-8

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences, 84*, 79-83. https://doi.org/10.1016/j.paid.2014.08.021

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

Minnesota Department of Education. (2024). *Minnesota Student Survey*. https://education.mn.gov/mde/dse/health/mss/

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867-872. https://doi.org/10.1016/j.jesp.2009.03.009

Rodriguez, M. C. (2023). *Technical report on developmental skills, supports, & challenges from the 2013-2022 Minnesota Student Surveys*. Minnesota Youth Development Research Group, University of Minnesota. https://hdl.handle.net/11299/255795

Scales, P. C., & Leffert, N. (2004). *Developmental assets: A synthesis of the scientific research* (2nd ed.). Search Institute. https://www.search-institute.org/product/developmental-assets-a-synthesis-of-the-scientific-research-on-adolescent-development/

Schneider, S., May, M., & Stone, A. A. (2018). Careless responding in internet-based quality of life assessments. *Quality of Life Research, 27*(4), 1077-1088. https://doi.org/10.1007/s11136-017-1767-2

Search Institute. (2013). *Developmental Assets Profile: Technical summary*. http://www.search-institute.org/surveys/dap

Search Institute. (2024). *Developmental assets*. https://www.search-institute.org/our-research/development-assets/

Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education, 32*(2), 151-165. https://doi.org/10.1080/08957347.2019.1577244

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, 38*(2), 101-111. https://doi.org/10.1111/emip.12256

van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement, 59*(4), 470–501. https://doi.org/10.1111/jedm.12317

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment, 28*(3), 189–194. https://doi.org/10.1007/s10862-005-9004-7