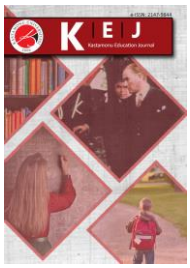
Kastamonu Education Journal, 2024, Vol. 32, No:3, 396-407
doi: 10.24106/kefdergi.1525320



|Research Article / Araştırma Makalesi|

Identifying the Careless Responders: A Cross-Country Comparison on PISA 2018 Dataset

Dikkatsiz Yanıtlayıcıların Belirlenmesi: PISA 2018 Verisinde Ülkeler Arası Bir Karşılaştırma[1]

Başak ERDEM-KARA[2]

Keywords
1.Careless Responding
2.Insufficient Effort Responding
3.Unmotivated Responding
4.PISA 2018
5.Self-report Data

Anahtar Kelimeler
1.Dikkatsiz Yanıtlama
2.Yetersiz Çabayla Yanıtlama
3.Motivasyonsuz Yanıtlama
4.PISA 2018
5.Öz-bildirim Verisi

Received/Başvuru Tarihi
18.04.2023

Accepted / Kabul Tarihi
11.07.2024

Abstract

Careless responding (CR) is one of the major concerns in self-report measures since it affects the quality of collected data. In this research, it was aimed to examine the dataset in terms of CR, to investigate the effects of CR on the psychometric properties of data, and to examine the effectiveness of CR detection methods in a cross-cultural context. Specifically; response time, outlier and long-string indices were used to flag CR and efficiency of using these were compared on PISA 2018 reading related attitude scales for Singapore and Turkey. As a result, it was indicated that the amount of CR was higher for Singapore than Turkey at long-string and response time methods. Furthermore, after removal of CR from dataset, enjoyment level of reading, perception of competence and perception of difficulty of PISA tests scores increased. Another result was that long-string and response time analyses provided significant improvement in model fit and long-string had the highest improvement level. Moreover, removing respondents according to outlier analyses didn't provide any improvement on model-data fit. In general, results illustrated that careless responding behavior may have important effects on psychometric properties of self-report and screening the data for CR is strongly suggested in future studies.

Öz

Dikkatsiz yanıtlama davranışı (DYD), öz-bildirime dayalı araçlarla toplanan verilerin niteliğini etkilemekte, bu nedenle bu araçları kullanan araştırmacılar için temel endişe kaynaklarından biri olmaktadır. Bu araştırma kapsamında, verileri dDYD bakımından incelemek, dikkatsiz yanıtlayıcıların verinin psikometrik özellikleri üzerindeki etkilerini araştırmak ve dikkatsiz yanıtlama davranışı gösteren bireyleri saptamada kullanılan yöntemlerin etkililiğini incelemek amaçlanmıştır. Bu doğrultuda; cevaplama süresi, uç değer ve uzun seri (log-string) yöntemleri kullanılarak dikkatsiz yanıtlayıcılar belirlenmiş ve bu yöntemleri etkililiği, Singapur ve Türkiye için PISA 2018 okumayla ilgili tutum ölçek verileri üzerinde karşılaştırılmıştır. Sonuç olarak, DYD miktarının Singapur için uzun dizi ve yanıt süresi yöntemlerine göre Türkiye'den daha yüksek olduğu belirtilmiştir. Ayrıca, dikkatsiz cevaplayıcıların veri setinden çıkarılmasının ardından, PISA testlerinden alınan okuma zevki, yeterlilik algısı ve zorluk algısı puanları artmıştır. Diğer bir sonuç ise, uzun dizi ve tepki süresi analizlerinin model uyumunda önemli iyileşme sağladığı ve uzun dizi yönteminin en yüksek iyileştirme düzeyine sahip olduğudur. Öte yandan, aykırı değer analizlerine göre yanıtlayanları işaretlemek ve veriden çıkarmak, model-veri uyumunda herhangi bir gelişme sağlamamıştır. Genel olarak, çalışma sonuçları dikkatsiz yanıtlama davranışlarının öz-bildirim verilerinin psikometrik özellikleri üzerinde önemli etkileri olabileceğini göstermiş ve ileriki çalışmalarda verinin bu bağlamda incelenmesinin önemi vurgulanmıştır.

[1] Part of the results of the study were presented at IX. International Eurasian Educational Research Congress by the researcher.

[2] **Corresponded Author,** Anadolu University, Faculty of Education, Eskisehir, Türkiye; https://orcid.org/0000-0003-3066-2892

Citation/Alıntı: Erdem-Kara, B. (2024). Identifying the Careless Responders: A Cross-Country Comparison on PISA 2018 Dataset, *Kastamonu Education Journal, 32*(3), 396-407. doi: 10.24106/kefdergi.1525320

INTRODUCTION

Self-report measures are widely used by researchers in many areas of social and psychological research. Those self-report instruments let researchers measure some psychological constructs (such as personality, beliefs, emotions, attitudes etc.) of lots of respondents in a short time (Alarcon & Lee, 2022; Curran, 2015; Ulitzsch et al., 2022). Ideally, it is assumed that respondents carefully read items and select the response reflecting themselves accurately by employing their maximal effort; so that scale scores truly reflect the measured construct (ability, proficiency, attitude etc.). However, this assumption is violated often since respondents, especially the unmotivated ones, may not willingly provide their maximum effort which is necessary to properly process a survey item and give accurate responses consistent with his/her latent trait (Rios & Soland, 2021; Schroeders et al., 2022; Ulitzsch et al., 2022). Datasets from self-report measures are prone to include several errors coming from inaccurate responding, inaccurate coding, or inaccurate computation (Huang et al., 2012). This kind of unmotivated response behaviors are labeled as random, careless, or inconsistent responding frequently in the literature (Alarcon & Lee, 2022).

Meade & Craig (2012) stated that inattentive/careless responding behavior is a major concern in any type of research based on surveys. Similarly, Brühlmann et al. (2020) identified respondents' carelessness as the major factor affecting the quality of data collected in surveys. Neglecting this responding behavior in self-report data, even the amount is small, may severely affect the study results and produce misleading findings since they may influence the measurement of the underlying factors in several ways (Alarcon & Lee, 2022; Bowling et al., 2021; Kountur, 2016; Ulitzsch et al., 2022). Rios et al. (2017) specified that low test-taking motivation of respondents may work as an external factor and make it difficult for responders to accurately demonstrate their skills, abilities or proficiencies being measured, endangering the validity of test-score interpretations. Beck et. al. (2019) stated that results may be affected at both item level and scale level because of inattentive responses which is a type of response bias. It may affect the results by introducing measurement error, weaking the relationship between variables and inflating the Type II errors. Construct-irrelevant variance is introduced to the measurement process and it can negatively influence the psychometric properties of the scale such as item difficulty, item discrimination, test reliability, factor structure etc. Shortly, inattentive/careless responses may potentially weaken the validity of test scores in several ways (Rios & Soland, 2021).

Recognizing the importance of careless responding, researchers and practitioners have shown a growing interest on reasons and effects of careless responses, how to detect and cope with them in order to guarantee the quality of data taken from surveys (Kountur, 2016). There are many different data screening techniques suggested to identify careless responders in the literature with its own advantage and disadvantage based on the situation (number of items, factor structure of the scale, number of subscales etc.). These methods will be discussed in detail later.

Karabatsos (2003) made a comparison of 36 person-fit indices concerning their accuracy of detecting aberrant responses on a dichotomous response scale over several simulated datasets. Person-fit statistics were evaluated with ROC analysis and results showed that in general, HT method performed best when detecting all types of aberrancies. Wise & Kong (2005), made a computer-based assessment and found that response time measure converged significantly with self-report measures and person-fit statistic index. In another study, Johnson (2005) stated that person-fit indices are too stringent, complex and computationally intensive. Meade & Craig (2012) examined different methods on an online questionnaire administration and a simulation. They examined bogus items, consistency indices, response pattern, outlier analysis, response time analysis and self-report measures. They stated that different methods flag different persons. There were two separate patterns of CR namely random and non-random and different indices were required to identify CR on these situations. As minimum screening tools, screening extremely short response times and self-report measure usage were recommended. As a result of simulation, they stated that Mahalanobis D can be an effective indicator of CR, but its power is too much dependent on properties of sample. Another study conducted by Huang et al. (2012) and they compared individual reliability, the psychometric antonyms index, the long-string and response times. They recommended three approaches namely response time, psychometric antonyms and individual reliability in future uses and stated that identification of effective CR methods may help researchers to improve their data quality. Goldammer et al. (2020) investigated the seven popular CR detection index and impact of CR on psychometric properties of constructs. They found that five of these seven indices namely response time per item, personal reliability, psychometric synonyms, psychometric antonyms, and Mahalanobis distance were effective on detection of CR. However, long-string and intra-individual response variability (IRV) were ineffective. In addition, when they examined the effect of CR, they realized that CR inflated item variances, biased item means towards the scale mid-point, increased residual variances of construct indicators. In addition to traditional methods, there are recently proposed more complex procedures suggested to identify CR. For instance, Schroeders et al. (2022) used one of the machine learning algorithms namely gradient boosted tree method (GBM) to identify careless responders and compared its performance with statistical outlier methods, consistency analyses, and response pattern functions. Both a simulation and an empirical study were conducted. As a result, it was shown that gradient boosting machines had performed better than traditional methods in simulation study. However, they stated that this performance of GBM did not transfer to the empirical study. Ulitzsch et al. (2022) proposed a new latent response mixture modelling approach stating that previous approaches are limited since they do not take person/item characteristics into consideration. They stated that traditional approaches identify CR at the aggregate respondent or scale level and they may fail to notice item properties evoking CR such as sentence length, item position, wording etc. Their proposed model allows for CR to vary at the item-by-respondent level. They indicated that their model provided a good parameter recovery and in a simulation data it was good at handling the simultaneous occurrence of multiple types of CR patterns. To examine the agreement of the proposed model and traditional methods, three frequently employed indicators namely

Mahalanobis distance, long-string index, the even-odd correlation, intra-individual response variability (IRV), and time spent on the questionnaire was used. They indicated that Mahalanobis distance and IRV methods had the strongest association with attentiveness, while correlations of long-string index and the even-odd correlation with the model were small. However, there was no substantive correlation with response time with neither attentiveness nor other indicators.

As previously mentioned, there were several studies on careless responding and its identification methods in the literature. However, there is still no clear answer concerning one of the most fundamental problems – detection accuracy of CR identification methods (Goldammer et al., 2020). Although, there are advantages of recently proposed methods such as machine learning algorithms or latent response mixture models. Machine learning algorithms have the disadvantage that they need an extensive training phase and prediction models are restricted to a particular set of items. Because every examination is particularly specific in terms of items and people it examines, generalizations to other data sets, samples, and situations are not conceivable (Schroeders et al., 2022). On the other hand, estimation with latent response mixture model could be quite computationally demanding and take a while (Ulitzsch et al., 2022). However, traditional methods suggested to identify CR have the advantage that they need minimal prerequisites.

In addition to the unanswered questions on CR identification methods, McFerran (2022) investigated CR as a function of respondent, survey and cultural characteristics and found that these characteristics may play a crucial role in careless responding. Grau et al. (2019) also reported that the amount of CR is expected to be higher for people in individualistic countries than in collectivistic countries. Goldhammer et. al (2016) also found that lower country test-taking effort is correlated with lower proficiency levels at the country level. Eklöf et al. (2014) studied on the TIMMS advanced data of Sweden, Slovenia and Norway and observed that Sweden had the lowest level of effort and achievement among those three countries. When students with low-level effort were excluded from dataset there were no significant difference in performance among countries although the difference were significant before the exclusion.

In the context of that study, some common and popular traditional CR detection methods were chosen and compared in cross-country context. This report is organized as follows. Firstly, a theoretical framework for careless responding and methods to identify CR is provided. Secondly, three popularly used and common identification techniques were chosen, studied and compared on a real dataset: PISA 2018 'reading-related attitudes' scale for Turkey and Singapore. Singapore was chosen since it was one of the best performer countries on PISA 2018. Turkey was one of the average performers. PISA scales were implemented in English for Singapore and in Turkish for Turkey sample. Lastly, results of the study were discussed with the related literature.

Careless responding: What is it and how to detect?

As explained before, respondents are expected to give their best while taking a survey; however, that is not the case at most of the time (Huang et al., 2012). Careless responding behaviour may occur when individuals respond to scale items independent of the actual item content. It can be observed when individuals respond without reading the item stem, misinterpret the item stem and/or answer options, be unmotivated to think about the item etc. (Huang et al., 2012; Ward & Meade, 2022). Different terms have been used in the literature to refer this behaviour: 'random response' (Beach, 1989), 'insufficient effort responding' (Huang et al., 2012), 'careless responding' (Meade & Craig, 2012), 'inattentive responding' (Beck et al., 2019), 'disengaged responses' (Soland et al., 2019), 'careless and insufficient effort responding' (Ulitzsch et al., 2022). In the context of this study, the term 'careless responding' with CR abbreviation is preferred.

The prevention of careless responding is not always possible and realistic. So, there is a need to identify careless respondents. Various data screening methods for identifying CR may be classified as priori and post hoc methods (Beck et al., 2019; Meade & Craig, 2012). Priori methods can be defined as the methods that are planned and included in a survey data collection process before the survey is administered. They are often based on some special items or scales assessing whether the respondent is paying attention to the item content or not (Beck et al., 2019). On the other hand, post hoc methods are implemented on collected survey data and they are typically based on calculating a statistic intended to detect aberrant response patterns (Beck et al., 2019; Meade & Craig, 2012). Some popular priori and post hoc methods are explained in the following sections.

A priori methods

Instructed response, bogus and self-report items can be stated as three of the most popular and effective a priori methods to detect CR (Beck et al., 2019; Schroeders et al., 2022; Ward & Meade, 2022). Instructed response items are items constructed to instruct the respondents to select a specific response category (e.g. "For this item, please select strongly disagree"). If respondents choose another response option, they are assumed as careless respondent. On the other hand, bogus items are items with an obvious answer (e.g. 'water is wet' (Gummer et al., 2021)). Respondents are required to be agree or disagree with that statement and failure to select the correct response option is considered as an indication of CR. Lastly, self-report items which are directly asking individuals about their engagement and effort (e.g. 'I put forth my best effort in responding to this survey'; Meade & Craig, 2012). Those items are generally placed at the end of the survey and sometimes known as validity check questions. Although these methods are straightforward and popular, the use of those items are still debateful since their involvement could have some negative spillover effects causing an irritation on participant (Curran, 2015).

Post-Hoc methods

Post-hoc methods are based on some special analyses after the completion of data collection process. There are several post-hoc methods to identify CR and most commonly used ones are consistency indices, response-pattern analysis, outlier analysis and response time analysis (Curran, 2015; Meade & Craig, 2012). Consistency indices use a variety of methods all of which are based on the idea that careless responders tend to provide internally inconsistent responses. For instance, one of the consistency indices is odd-even consistency which is based on the within person correlation across unidimensional subscales formed by splitting up each respondents' responses into even-odd item sets (Schroeders et al., 2022). However, it is not too effective for scales that do not include many items and scales (Niessen et al., 2016). In the context of response-pattern analysis, response pattern functions can be used to identify unusual patterns in either parts or entire response vector of an individual compared to the others. Most commonly used indices are long-string index and intraindividual response variability (IRV). Long-string index are computed by looking at number of times a respondent gave the same response option in a row. Too much choice of a single response option is thought to be an indicator of CR (Meade & Craig, 2012). On the other hand, IRV calculates the inter-item standard deviation across a response vector (Meade & Craig, 2012; Schroeders et al., 2022). There are some more complex procedures of response pattern functions such as person-fit statistics (PFS). PFS examine the probability of getting the observed pattern of responses and can be used to identify aberrant response patterns that deviate significantly from expected ones (Meijer, Niessen, & Tendeiro, 2016; Meijer & Sijtsma, 2001). Poor PFS come up when respondents agree with items that are more extreme while disagreeing with those that are similar to or less extreme. Several indices can be computed and examined in detail from Beck et al. (2019)'s work. Furthermore, the identification and filtering of statistical outliers is defined as one of the most commonly used methods for data cleaning. Multivariate analysis such as Mahalanobis distance consider the entire response patterns across a series of items (e.g. a scale) so they are more appropriate to identify extreme cases than the univariate outlier analysis. Lastly, response time analysis can be used for detection of CR. Response time is the time that an individual spends to respond a set of items and Curran (2015) stated it is perhaps the most commonly used method to eliminate the CR. It is not common that respondents read the item and give her/his reaction seriously when the response time is too short and very short response time is assumed to be an indicator of CR. Huang et al. (2012) suggested a cutoff score for response time that is, a response to an item could be given in more than 2 seconds. On the other hand, Zhang & Conrad (2014) suggested another cut score based on the number of words in the item and stated that 300 msec is required for each word in the item. For instance, if there are 5 words in the item, respondents answering that item less than 1.5 sec are identified as careless.

As can be seen, there are several methods to detect CR and some of most popular ones were explained above. For the detailed review of other methods, you can see Curran (2015) and Meade & Craig (2012). In the context of current study, a priori methods couldn't be used since the used PISA 2018 dataset was taken from OECD website. It is ready to use dataset and researchers have no control on it. Among post-hoc methods; response time, outlier and long-string analysis were used as CR detection methods in this research.

Present study

The main purpose of this study is to evaluate the effects of careless responses identified by different techniques on PISA 2018 scales with a cross-country comparison. Therefore, three of the popular techniques namely response time, long-string and outlier have been used to identify CR among Turkish and Singaporean students on PISA 2018 in the first stage. In the context of this research, answers were sought for the following research questions:

What were the descriptive characteristics of Singaporean and Turkish students on the main dataset?

How was the distribution of careless respondents in Singapore and Turkey according to three different methods?

When respondents were removed from dataset, what kind of changes were observed?

After deleting flagged careless responders, how much improvement in model fit was obtained?

METHOD

Data

Data used in this study were PISA 2018 Student Questionnaire items focusing on measuring reading-related attitudes and data was obtained from OECD website. Due to the nature of the study, informed consent or ethics committee approval was not required.

There were four indicators focusing on individual's reading-related attitudes in total namely; enjoyment of reading, self-concept of reading in two aspects (perception of competence and perception of difficulty) and perception of difficulty of the PISA test. Each of these indicators were investigated separately considering each one of them as a scale in PISA context. The general demonstration and number of items on these fields is summarized in Figure 1 below.

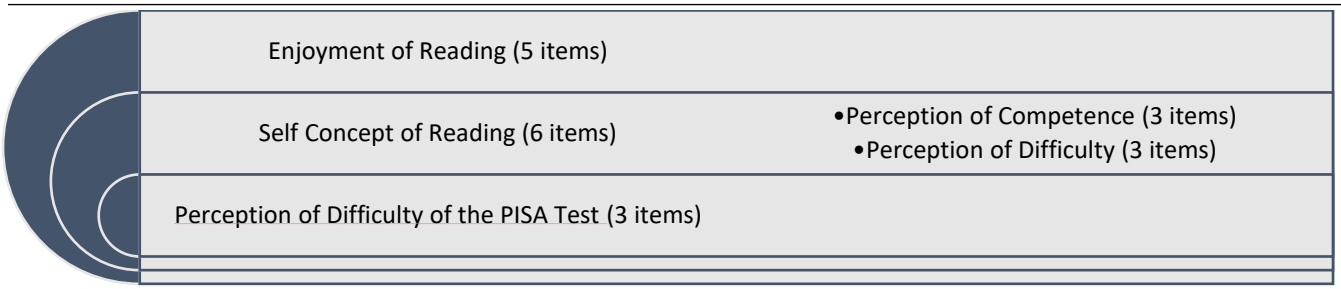


Figure 1. PISA 2018 reading-related attitudes scales

As can be seen in Figure 1, there were 14 items in total each of which have four response categories going through "Strongly disagree", "Disagree", "Agree", to "Strongly agree". There were three items which negatively worded and all of them were on 'Enjoyment of Reading' scale. These items were reverse coded such that higher scores on that scale imply higher enjoyment levels. Scale items were presented in the appendices part in English version.

There were 6676 students from Singapore and 6890 students from Turkey participated in PISA 2018 in total. Before the analysis, dataset was checked for missing values and individuals with missing responses were excluded from dataset by listwise deletion. After removing missing data, analysis continued with 6434 students from Singapore and 6387 from Turkey. 745 students in total (503 from Turkey (7.3%) and 242 from Singapore (3.76%)) were deleted from the dataset.

Careless responding analyses

Response time index. Firstly, response time method was used to identify CR. Response time data is included in the PISA 2018 dataset at OECD website in the file named as 'Questionnaire timing data files'. In the context of PISA, data file includes response time demonstrating the duration that a respondent spends to finish a scale on the related page. So, it can be considered as page time and did not provide an item-level data. For instance, since there are four separate scale presented on four separate pages, there are four separate response time data for each respondent on the file.

In different studies, total scale time indicating the duration between a participant begins and finishes the questionnaire was used as a RT measure. However, there are some debates on the use of total time and some prior studies indicated that total completion time demonstrated poor convergent validity with other CR techniques (Beck et al., 2019). Another index named as 'page time' was proposed by Huang et al. (2012) as an alternative and used in different studies. It is used for multi-page questionnaires and computed by taking the completion time of the scale on that page for each respondent. In the context of that study, 'page time' index was calculated by using 2 seconds rule for each item (Huang et al., 2012). After completing those procedures for each questionnaire on different pages, each record was averaged and 'page time index' was obtained.

Long-string analysis. Another method used to identify careless respondents was long-string analysis that was based on the examination of the longest string of identical responses of each individual. In this technique, it is assumed that careless respondents may have potential to choose the same response option to every question. The ones who are responding carefully and with sufficient effort will not select the same option for long periods of times (Curran, 2015). After calculation of longest string values, a cut-off should be carefully determined to determine careless respondents. Curran (2015) suggested a baseline rule that if respondents have a string of identical responses on at least half of the test (greater than or equal to the scale length), they are considered as careless respondents.

Outlier analysis (Mahalanobis distance). Mahalanobis distance across the entire scale was computed in order to identify multivariate outliers and outliers were flagged as CR.

All analyses were conducted in RStudio environment using the careless (Yentes & Wilhelm, 2021) package for outlier and long-string analyses and lavaan (Rosseel, 2011) package for factor analyses. Other parts in the analysis were conducted with the codes written by the researcher.

FINDINGS

Descriptive statistics

Table 1 indicates descriptive statistics for scales on Singapore and Turkey sample. Reverse items (Items 1, 4 and 5 on Scale 1) were recoded such that higher scores mean higher enjoyment of reading. Besides, as indicated above, missing values were excluded from dataset before the analysis. Values on Table 1 are based on the recoded data and remaining students on dataset after removal of missing values.

Table 1. Descriptive Statistics on PISA 2018 Reading Related Attitudes Scales

		S1	S2	S3	S4	Total
Turkey	M	14.92	8.65	6.06	5.80	35.496
	Sd	3.37	1.96	1.85	1.85	4.61
	α	0.80	0.80	0.69	0.70	0.76
Singapore	M	13.01	8.36	6.78	5.79	34.06
	Sd	3.84	1.97	1.85	2.08	4.63
	α	0.88	0.82	0.72	0.87	0.81

M:Mean Sd: Standard Deviation α: Cronbach alpha

As shown in Table 1, Turkish students had higher score on average on S1 and S2 subscales than students in Singapore and the difference of scores were significant (p<.05) for both subscales. It means that Turkish students' enjoyment of reading and perception of competence was significantly higher than Singaporean students. The only scale that Singaporean students had higher mean value was S3 scale (p<.05) and it was about the perception of difficulty. S4 scores that were related to perception of difficulty of PISA tests were very close to each other (TUR: 5.80, SNG: 5.79; p > .05). In addition to that, reliability of all subscales was calculated by Cronbach alpha coefficient and greater than 0.69 indicating adequate reliability. Specifically, Singapore data was more reliable than Turkey data on all scales and the difference between reliability coefficients were statistically significant (p < .05) through all scales. Besides, stratified alpha coefficient was also calculated for general scale and it was 0.757 for Turkey and 0.81 for Singapore (p < .05). It can be said that data taken from Singaporean students was more reliable than Turkish sample both on subscale and general scale level.

Identification of careless respondents

Response Time Analyses. In CR identification based on response time method, threshold value was set on 10000, 12000 and 6000 milliseconds for enjoyment of reading, self-concept of reading and perception of difficulty of the PISA test scales respectively. These thresholds were defined according to the number of items at scale so as to be two seconds for each item. It should be noted that response time measures were taken as the time that respondent spent to complete a scale on one page, not the entire survey, and there was no item-level information on data file. Since response time data in PISA files were given in milliseconds, our thresholds were also given as milliseconds. According to the response time analysis, page index was produced for each country. As can be seen in Table 2 below, 28 students from Turkey and 45 from Singapore spent less than specified cut off values and were flagged as 'careless' (and coded as 0) on all three scales (Page Index=0). On the other hand, 0.33 value of page index indicated that students had spent enough time only on one of the scales and flagged as 'careless' on two of the scales. In the current study, 0.33 was accepted as the page index cut off score to flag individuals as 'careless'. That is, there were 99 (1.6%) careless respondents from Turkey and 156 students (2.4%) from Singapore according to response time analysis.

Long-String Analyses. Lastly, long-string analyses were conducted by setting the cut-off value as 7 since there were 14 items in total. Long-string analyses were made on the original (non-recoded) data since data recoding changes the response pattern. The number of respondents with strings is given in Figure 2 below.

Figure 2 indicated that the longest string was 3 for both countries and it was observed at 2306 respondents from Turkey and 2344 from Singapore. As a cut-off, seven was identified since the scale has 14 items. As a result, it was observed that in total there were 341 and 384 individuals who had a string equal to or greater than seven from Turkey and Singapore respectively and those were coded as 'careless'.

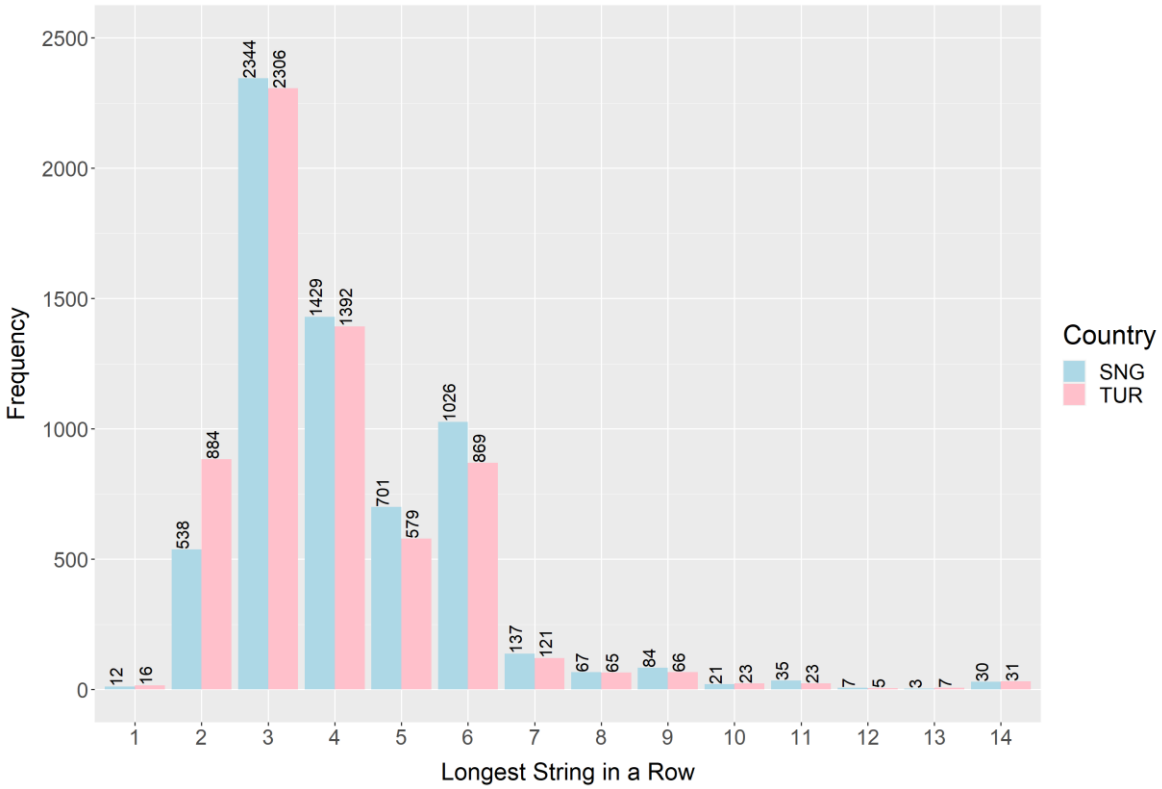


Figure 2. Number of responders with longest string

Outlier Analysis. Multivariate outliers were identified by computing Mahalanobis distances on all items on the scale and flagged as CR. According to the outlier analysis, there were 328 individuals from Turkey and 242 from Singapore required to be excluded from dataset.

Table 2 shows the number of careless respondents at Turkish and Singaporean students according to three different methods.

Table 2. Number of careless responders according to three methods

	Page Index			Outlier	Long-String Index
	0	0.33	Total		
Turkey	28	71	99 (1.6%)	328 (5.1%)	341 (5.3%)
Singapure	45	111	156 (2.4%)	242 (3.8%)	384 (6%)

The second research question was related to the prevalence of careless responding among Turkish and Singaporean students from PISA 2018 reading related attitudes scale. Estimates ranged from 1.6% to 5.3% depending on which IER detection method was applied. The method flagged the largest number of students as careless is long-string for both Turkish and Singapore sample. On the other hand, page index that was computed based on response time data had the lowest number of flagged students. Detection rates were lower on Turkish students' data than in Singapore sample for page-index and long-string but for the outlier analyses it was vice versa. CR rates on two countries were actually not too far from each other for three methods.

In Table 3, the correlations of the students flagged by three different CR methods were presented. Tetrachoric correlation was used for computation since our variables were binary.

Table 3. Correlation between three CR identification methods

	TUR			SNG		
	Response Time	Outlier	Long-String Index	Response Time	Outlier	Long-String Index
Response Time	1.00	0.40	0.59	1.00	0.43	0.59
Outlier		1.00	0.24		1.00	0.14
Long-String Index			1.00			1.00

From Table 3, it can be concluded that correlations between methods were highest between response time and long-string index for both countries (r_{TUR}= 0.59, r_{SNG}=0.59). On the other hand, the lowest correlation was between outlier and long-string methods (r_{TUR}= 0.24, r_{SNG}=0.14). It can be said that correlation between response time-outlier and response time-long string index was in moderate level. Correlation between outlier and long-string methods was in a low level.

Exclusion of Careless Respondents

With CR analyses, careless responders were flagged based on three different methods. After that, flagged students were removed from dataset according to the findings of three CR detection methods. Namely, three different datasets were formed. Descriptive statistics of remaining datasets were presented on Table 4.

Table 4. Descriptive Statistics after Exclusion of CR Flagged Students

			Remaining Dataset				
			S1	S2	S3	S4	Total
Turkey	*Response Time*	*M*	14.95	8.67	6.03	5.86	35.51
		Sd	3.38	1.94	1.82	1.83	4.56
		α	0.80	0.79	0.68	0.70	0.76
	Outlier	M	15.05	8.68	5.99	5.81	35.53
		Sd	3.31	1.89	1.74	1.78	4.46
		α	0.82	0.81	0.69	0.71	0.77
	Long-String Index	*M*	15.02	8.70	5.98	5.81	35.52
		Sd	3.39	1.92	1.78	1.81	4.47
		α	0.81	0.78	0.66	0.68	0.74
Singapore	Response Time	M	13.01	8.35	6.76	5.89	34.02
		Sd	3.86	1.95	1.82	2.06	4.6
		α	0.88	0.81	0.71	0.86	0.81
	Outlier	*M*	13.09	8.39	6.75	5.86	34.08
		Sd	3.84	1.91	1.79	2.03	4.49
		α	0.89	0.82	0.72	0.88	0.82
	Long-String Index	M	13.08	8.41	6.73	5.84	34.06
		Sd	3.91	1.97	1.84	2.07	4.60
		α	0.88	0.81	0.71	0.87	0.81

When Table 4 was examined, it was realized that with deletion of CR from dataset total scale scores increased a little compared to the whole dataset for Turkey with respect to each of three methods. Similarly, for Singapore, total scale scores did not differ between original dataset and CR excluded dataset with respect to long-string index, increased a little by outlier method and decreased with response time. However, there were no big differences with the main data set. Total scale scores of main datasets were 35.50 for Turkey and 34.06 for Singapore (Table 1). Comparing excluded dataset S1 values with original dataset indicated that mean value of S1, S2 and S4 scores increased with exclusion of CR at all methods. It can be interpreted that careful responders had higher enjoyment of reading, perception of competence and perception of difficulty of PISA tests scores compared to the excluded ones flagged as careless. On the contrary, S3 scores decreased with exclusion of CR at all methods which meant that careful responders had lower score of perception of difficulty compared to careless ones. Reliability values were also compared between excluded and original datasets. Cronbach alpha values were calculated for each scale separately and a stratified alpha value was calculated for whole scale. It was seen that obtained reliability coefficients on excluded datasets was not too far from their counterparts on the original dataset. The noticed point was that exclusion of respondents with respect to long-string index reflected on the dataset negatively in terms of reliability for Turkey dataset. Examinations can be done in more detail by comparing Table 1 and Table 4.

CFA Analyses

After the examination of descriptive statistics, model-data fit was investigated with confirmatory factor analyses (CFA) on remaining dataset according to three methods. In order to investigate the model-fit change between data before and after exclusion of CRs, CFA fit indices were used. CFA analyses were conducted on RStudio with 'lavaan' package with WLSMV estimator. Since all scale items have four response categories and the multivariate normality wasn't met, WLSMV was preferred. Fit indices are reported in Table 5 below.

Firstly, when fit indices of baseline model were investigated it was seen that the model-data fit is in an acceptable level for both Turkey and Singapore (CFI>.90, TLI>.90, RMSEA<.08, SRMR<.08). Singapore had better model data-fit since its' CFI and TLI values were higher and SRMR was lower compared to Turkey. As can be seen, removing participants according to response time and long-string methods increased CFI and TLI scores and decreased RMSE, SRMR values which means that it offered a significant improvement in model-data fit (ΔCFI <. -01) for both countries. According to CFI, TLI, RMSE and SRMR values, long-string method was the one providing the most significant improvement between raw dataset and the remained dataset. Removing respondents who were flagged according to response time index also significantly improved the model fit. However, outlier method didn't provide any improvement on fit indices in comparison with raw data.

Table 5. Fit indices of original (baseline) and remained datasets

	Method	CFI	TLI	RMSEA	SRMR
SNG					
	Baseline	0.935	0.917	0.060	0.040
	Response Time	0.937	0.919	0.060	0.039
	Outlier	0.928	0.907	0.065	0.039
	LongString	0.939	0.921	0.060	0.038
TUR					
	Baseline	0.909	0.883	0.060	0.048
	Response Time	0.913	0.889	0.059	0.046
	Outlier	0.907	0.881	0.065	0.046
	LongString	0.925	0.903	0.056	0.042

DISCUSSION

Although importance of possible problems caused by careless responding (CR) in survey data have been emphasized in several studies, data screening for CR is not common practice. This lack of screening might be due to the fact that although there are several studies on careless responding and its identification methods in the literature, there has been still no clear answer concerning the effectivity of CR identification methods. As a contribution to that problem, the effect of careless respondents flagged by three of popular methods namely response time, outlier and long-string indices to identify CR were examined and the performance of three identification methods were compared on a real dataset. Some psychometric properties of the scale and model-fit was compared after removing flagged CRs.

Consistent with previous researches (Goldammer et al., 2020; Huang et al., 2012; Meade & Craig, 2012; Ulitzsch et al., 2022; Wise & Kong, 2005), as a result of this research, response time was determined as an effective method to identify CR. It provided significant improvement in model fit. However, in contrast to previous findings (Goldammer et al., 2020; Huang et al., 2012), long-string index was stated as the one providing highest model-data fit improvement so it was taken as the most effective strategy to identify CR as a result of this research. As Johnson (2005) stated, long-string analyses is based on the assumption that if a person consistently gives the same response, his/her reactions may not be responsive to the item content. It assumes that there should be a variability in responses of a person. This technique tends to be affected by scale or sample properties. That is why it is hard to compare findings from long-string analyses across different data collections (Curran, 2015; Johnson, 2005). Curran (2015) stated that long-string analysis provide opportunity to eliminate some of the worst of the worst responders, but it might be challenging to accomplish much more. Response time and long-string methods are suggested as a bare minimum for the removal of careless responders and a good start before implementing more advanced techniques (Curran, 2015). Despite of the limitations that long-string method has, it is better to exclude worst responders than doing nothing at all. In addition to that, because of its' scale and sample-specific nature, there is no global cut-off score to identify careless ones and it can be decided in different ways in different studies which can result in different findings.

Another result of this study was that using outlier analyses didn't provide any improvement and that finding was not in line with previous studies findings (Goldammer et al., 2020; Meade & Craig, 2012; Ulitzsch et al., 2022). Meade & Craig (2012) stated that this technique may be affected by both deviations from normality in items and too much normality in careless responders. Besides, careless responders could be well versed at giving their responses that are close to the midpoint of other responses and could not be flagged by this method. It was expected that different CR detection methods flag different responders and it is quite understandable. For instance, the long-string index won't be able to identify someone who answers questions at random, but another technique might.

When descriptive statistics were investigated in detail, it was realized that total scale score increased with exclusion of careless responders for Turkey but not for Singapore. But, examining scales separately in detail, it was indicated that S1, S2 and S4 scores increased and S3 scores decreased with exclusion of CR at all methods. It means that careful responders had higher level of enjoyment of reading, perception of competence and perception of difficulty of PISA tests scores and lower score of perception of difficulty of reading than excluded responders who were flagged as careless. It has frequently been reported that unmotivated participants frequently provide careless responses (Rios & Soland, 2021; Schroeders et al., 2022; Ulitzsch et al., 2022; Wise & Kong, 2005). Mol and Bus (2011) stated that since motivation influences how much and how widely kids read, which in turn helps students develop their reading competency, motivation is thought to be especially important. Students with high interest are more likely to be engaged in reading tasks (Grabe, 2009). So, it was expected that level of enjoyment of reading and perception of competence of careful responders were higher and perception of difficulty of reading were lower than careless ones. About the perception of difficulty of PISA tests, it can be considered that since careful responders were assumed to give more effort for the test and they were more likely to be engaged, they may have perceived the test more difficult. In addition, Cronbach alpha for separate scales and stratified alpha for general scale was computed to observe possible changes in alpha after exclusion of flagged CR. The results indicated that removing CR did not led to big differences on reliability coefficient. The biggest difference was on that exclusion of respondents with respect to long-string index reflected on the dataset negatively in terms of reliability for Turkey dataset. Wise & DeMars (2006) worked on a cognitive test and observed that after removing low-effort students, coefficient alpha

of the test dropped to 0.66 from 0.84. Similarly, Guo et. al. (2016) reported that exclusion of rapid guessers led to a decrease in reliability coefficient. So, the result was not unexpected.

When findings of the study were examined in country level in detail, it was discovered rate of careless responders were higher in Singapore sample for response time and long-string index and lower for outlier method. Since outlier method was the one providing no improvement at model-data fit after removal of careless responders according to CFA results, response time and long-string index were taken into consideration for interpretation. As Grau et al. (2019) stated, some differences in CR among countries could be expected. They mentioned that the amount of CR is expected to be lower in individualistic countries than in collectivistic ones since people who live in individualistic countries may be more accustomed to being concerned with personal attitudes, feelings or evaluations. They may also be more likely to respond truthfully to questions about their personality and attitudes when asked in a questionnaire. When the Individualism Distance Index (IDV) of Turkey and Singapore were examined, it was seen that Turkey had higher IDV than Singapore which means that it is more individualistic. So, it was expected that the amount of CR was lower in Turkey than in Singapore.

IMPLICATIONS AND RECOMMENDATIONS

As results illustrated, careless responding may have important effects on psychometric properties of self-report and screening the data for careless responding, which is actually an extension of more conventional data cleaning, is strongly suggested in future studies. The findings of this study may offer practitioners recommendations on how to deal with careless responding especially on self-report scales. Considering how widely self-report scales are used in the field of educational research, the findings are important to increase educational researchers' awareness on careless responding and how to deal with it. Besides, present study adds to the literature by focusing on cross-cultural comparison on a real data set but there are still unanswered questions.

The results of this study are limited with the particular sample and measures used. Especially long-string method is a sample and scale-specific method and may give different results in different conditions. As another method, response time could not be examined in item level since data were provided only at scale level. That is why, page time index was computed and evaluated. As another limitation, the number of items on scales was not too much, it was a 14-item scale.

Future research may consider to use different CR identification methods on different samples and longer scales. Since data used in this research is ready to use and researcher has no control on it, post-hoc methods were used and the accuracy or sensitivity of the methods could not be examined in detail. Future research may consider to collect data and take priori methods into consideration. Besides, some personal characteristics such as gender, age, motivation could also be examined on how they affected CR.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, author-ship, and/or publication of this article.

Statements of publication ethics

I/We hereby declare that the study has not unethical issues and that research and publication ethics have been observed carefully.

Researchers' contribution rate

The study was conducted and reported by the researcher herself on each stage.

Ethics Committee Approval Information

In the context of this study, PISA 2018 dataset which can be directly downloaded from OECD website by everyone was used. It was a ready-to-use dataset and there was no data collection process carried out by the researcher. So, ethics committee approval was not required.

REFERENCES

Alarcon, G. M., & Lee, M. A. (2022). The relationship of insufficient effort responding and response styles: An online experiment. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.784375

Beach, D. A. (1989) Identifying the random responder. *The Journal of Psychology*, *123*(1), 101-103, DOI: 10.1080/00223980.1989.10542966

Beck, M. F., Albano, A. D., & Smith, W. M. (2019). Person-fit as an index of inattentive responding: A comparison of methods using polytomous survey Data. *Applied Psychological Measurement*, *43*(5), 374–387. https://doi.org/10.1177/0146621618798666

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2021). The quick and the careless: the construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*. https://doi.org/10.1177/10944281211056520

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*. https://doi.org/10.1016/j.metip.2020.100022

Curran, P. G. (2015). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*(2016), 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, *27*(1), 31–45. https://doi.org/10.1080/08957347.2013.853070

Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32* (3), 224–247. https://doi.org/10.1177/0146621607302479

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

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). Test-taking engagement in PIAAC. Vol. 133. In: OECD Education Working Papers. Paris: OECD Publishing.

Grabe, W. (2009) Reading in a second language: Moving from theory to practice. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139150484

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural differences in careless responding. *Journal of Cross-Cultural Psychology*, *50*(3), 336–357. https://doi.org/10.1177/0022022119827379

Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, *50*(1), 238–264. https://doi.org/10.1177/0049124118769083

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J. & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education*, *29*(3), 173-183, doi: 10.1080/08957347.2016.1171766

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*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39, 103–129. http://dx.doi.org/10.1016/j.jrp.2004.09.009.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16* (4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Kountur, R. (2016). Detecting careless responses to self-reported questionnaires. *Eurasian Journal of Educational Research*, *16*(64), 1–35. https://doi.org/10.14689/ejer.2016.64.17

McFerran, M. W. (2022). *Careless responding in survey research: an examination of individual, situational, and cultural characteristics* (Doctoral dissertation). Florida Tech.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A practical guide to check the consistency of item response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment*, *23*(1), 52-62. https://doi.org/10.1177/10731911155778

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107-135. https://doi.org/10.1177/0146621012203

Mol S. E. & Bus A. G. (2011) To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*(2), 267–296. doi:10.1037/a0021890.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, *17*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

Rios, J. A., & Soland, J. (2021). Parameter estimation accuracy of the effort-moderated item response theory model under multiple assumption violations. *Educational and Psychological Measurement*. http://journals.sagepub.com/doi/10.1177/0013164420949896

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, *82*(1), 29–56. https://doi.org/10.1177/00131644211004708

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

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. Journal of Statistical Software, *74*(5), 1–27. https://doi.org/10.18637/jss.v074.i05

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3). https://doi.org/10.1111/bmsp.12272

Ward, M. K., & Meade, A. W. (2022). Dealing with careless responding in survey data: prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*(1). https://doi.org/10.1146/annurev-psych-040422-045007

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Yentes R.D., & Wilhelm, F. (2021). Careless: Procedures for computing indices of careless responding. R package version 1.2.1. https://cran.r-project.org/web/packages/careless

Zhang C. & Conrad F. G. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453