Survey Research Methods (2021)
Vol. 15, No. 3, pp. 299-325
doi:10.18148/srm/2021.v15i3.7832
European Survey Research Association

The Response Entropy Index: Comparative Assessment of Performance and Cultural Bias across Indices of Careless Responding

John Tawa
Mount Holyoke College
South Hadley, U.S.A.

The response entropy (RE) index is proposed as a new method for flagging careless response patterns and is determined by calculating the balance of proportions of response types endorsed by participants on Likert-scaled surveys. In the first study, performance of the RE index was compared to other commonly used post hoc indices for detecting careless responding (CR) such as the Mahalanobis distance (MD) and the psychometric synonym (PS) index. Three different types of Bogus Sets (BS) were generated: 1) uniform random values produced by computer (n = 100); 2) normally distributed random values produced by computer (n = 100); and 3) purposefully careless responses produced by human participants (n = 100). The BS data were then implanted in a true, cleaned social science dataset (n = 500). Multinomial logistic regression determined that the RE index made independent contributions from other indices to the prediction of BS. Latent variable analyses suggest that the variability type RE index may be tapping distinct constructs from regression type indices such as the PS index. In study 2, potential cultural bias in CR indices was examined with a true social science dataset (n = 302) comprised of racially diverse participants. Unlike other post hoc indices of CR, the RE index was unrelated to participant race. Further analyses demonstrated that racial differences on other indices of CR could be accounted for by culturally different styles of survey responding. For example, Asian participants' higher MD scores relative to White participants' was mediated by a culturally specific acquiescent survey response style. These findings point to the useful of the RE index for detecting CR while also avoiding the conflation of CR with culturally different responding.

Keywords: careless responding; cultural sensitivity

1 The Response Entropy Index: Comparative Assessment of Performance and Cultural Bias across Indices of Careless Responding

Today, researchers increasingly use online survey hosting sites to collect survey data. While there are many benefits to using online servers, including ease of administration and data management, one challenge is that participants may be more likely to respond carelessly to online survey items compared to paper and pencil surveys which may, for example, be taken in the presence of an administrator. Perhaps online responders feel less responsibility to answer attentively because of their anonymity (Johnson, 2005). Regardless of the reason, careless responding affects the quality of data and increases the chances of Type II error (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Marjanovic, Holden, Struthers, Cribbie, & Greenglass, 2015), and in some cases may even inflate Type I error rates (due to careless responders clustering around mid-point responses; Huang, Liu, and Bowling (2015)). The prevalence of careless responding is difficult to estimate and depends considerably on the method used for estimation. Clinical and personality measures such as the MMPI (Minnesota Multiphasic Personality Inventory) and NEO (Neuroticism, Extroversion, Openness Inventory) have embedded careless response indices such as the Variable Response Inconsistency (VRIN) subscale, a pre-determined set of similar or opposite-valanced item pairs that when responded to inconsistently indicate a subject who is not attending to content (Berry et al., 1992; Johnson, 2005; D. S. Kim, McCabe, Yamasaki, Louie, & King, 2018; Kurtz & Parrish, 2001). Prevalence estimates made on these bases range from 3.5% (Johnson, 2005) to 10.6% (Kurtz & Parrish, 2001). Self-report methods involve asking participants directly the approximate proportions of questions on the test to which they were "unable to pay attention to and answered randomly" (Berry et al., 1992, p. 341). Although self-report methods do positively correlate with embedded random response indices (e.g. Berry et al., 1992), prevalence estimates are considerably higher and range from as many as 50%

Correspondence address: John Tawa, Department of Psychology and Education, Mount Holyoke College, 50 College St., South Hadley, MA, 01075 (E-mail: jtawa@mtholyoke.edu).

(Berry et al., 1992) to 73% of participants indicating randomly responding to one or more items (Baer, Ballenger, Berry, & Wetter, 1997). Although question remains about the severity of the problem, advances in methods for detecting careless responding should nonetheless contribute to improving data quality for researchers using survey designs.

In this paper, I offer a new index for detecting careless response patterns, the response entropy (RE) index. I am particularly concerned with detecting careless responding (CR), meaning participant response patterns that are hurried, answered without comprehension of the item, and do not reflect participants' true feelings or thoughts related to the items. CR should be differentiated from other forms of inaccurate responding such as intentionally answering items inaccurately. For example, some participants may "fake good" or, respond in a way to make oneself appear more well-adjusted or socially desirable than is actually the case (Crowne & Marlowe, 1960; Roma et al., 2019). Intentionally answering incorrectly suggests that the participant has read and understood the item but for some reason is motivated to answer incorrectly (Huang et al., 2012; Johnson, 2005; Meade & Craig, 2012). In contrast, CR has been referred to as "content independent" (Huang et al., 2012) or "content nonresponsivity" (Meade & Craig, 2012)) and is indicative of people who respond to items while only scanning the questions or not reading them at all. The RE index is intended for detecting CR only, and not for detection of other forms of inaccurate responding.

The problem of careless responding on surveys, whether online or on paper, is not a new challenge for researchers. Researchers in the past have used at least four approaches to address the problem of CR. The first approach has been to explicitly deter CR, for example, by including instructions or warnings on the survey about carelessly responding to items (Huang et al., 2012; Meade & Craig, 2012). The second approach was discussed earlier and involves including a predetermined set of similar or opposite-valanced item pairs in the survey that should be answered to consistently (Berry et al., 1992; Johnson, 2005; Kurtz & Parrish, 2001). An alternative to this approach is to include one or more improbable items within the survey; for example, Beach (1989) recommends the inclusion of the items (e.g., "I was born on February 30th") to which endorsement would suggest participants are not attending to the content of items. The third approach has been to assess participants' response times for answering survey questions. Rapid responding has been found to be associated with self-reported carelessness in responding (Huang et al., 2012; Leiner, 2019; Wise & Kong, 2005). Huang et al. (2012) operationalized response time as the average amount of time participants spent on each survey page on an online survey. The fourth and final approach is the one adopted in this paper and differs from the previous three in that it is a *post hoc* approach (Marjanovic et al., 2015); it does not require previous inclusion of items or assessments (such as configuring surveys to record response time) and can be used with any existing data set. This approach involves the computation of indices of CR that are calculated based on participants' response patterns across a survey. Below, I first review previous post hoc methods of assessing CR, and then discuss the concept of the response entropy index.

1.1 Post Hoc Methods of Assessing CR

The development of computational methods for detection of CR from survey response patterns is relatively recent and has received increased attention in the past decade. Here I propose that these indices generally fall within three classes: regression type, person-fit type, and variability type. The response entropy index may be most accurately classified as a variability type index.

Regression type indices work on the general assumption that purposeful responders will demonstrate a pattern of scores on items within a survey that are relatively consistent, or *correlated* with one another (Huang et al., 2012; Johnson, 2005; Meade & Craig, 2012). Computation of regression type indices generally involves determining within-person correlations among a specified subset of items within a survey. For example, in the even-odd index even and odd numbered survey items are matched and a within-person correlation is conducted; higher values indicate more consistency within the responder and by proxy, less carelessness. There are at least three such indices offered in the literature: the even-odd (EO) index (Johnson, 2005), the psychometric antonym (PA) index (Goldberg & Kilkowski, 1985), and the psychometric synonym (PS) index (Meade & Craig, 2012). The Mahalanobis distance (MD) is a multivariate outlier analysis and determines the distance of a single person's score to the overall covariance pattern of the sample (Mahalanobis, 1936; Meade & Craig, 2012; Niessen, Meijer, & Tendeiro, 2016). Each of these are described in more detail and operationalized in the methods section.

Person-fit type indices are based on item response theory (IRT; Böckenholt (2013), Lang, Lievens, De Fruyt, Zettler, and Tackett (2019), Sijtsma and Molenaar (2002)) and weigh the relative probability of responses to each item rather than assume that responses to each item are equally probable. Person-fit indices, then, provide a measure of the extent to which a specific participant's pattern of item responses fit with the overall sample's pattern of responses. For example, imagine that the probabilities of item responses (based on the overall sample) for the first two items on a five point Likert-scale index are [0.1, 0.3, 0.2, 0.2, 0.2] and [0.3, 0.2, 0.2, 0.2, 0.1]; a respondent who answers 2 and 1 respectively demonstrates the highest possible fit, whereas a respondent who answers 1 and 5 respectively demonstrates the lowest possible fit. Specific person-fit type indices have been used for detection of CR, including the number of Guttman errors

(Niessen et al., 2016; Schneider, May, & Stone, 2017; Meijer, Niessen, & Tendeiro, 2016) and the standardized log-likelihood (Niessen et al., 2016; Meijer et al., 2016; Drasgow, Levine, & Williams, 1985) of a response pattern under an IRT model.

Variability type indices work on the general assumption that both too much and too little variability in a response set are characteristic of careless responding (Costa Jr & McCrae, 2008; Huang et al., 2012; Lang et al., 2019; Leiner, 2019; Marjanovic et al., 2015). Purposeful item responses tend to cluster around a personal mean score. For example, a person with relatively severe depression may respond on a six point Likert-scaled depression inventory with primarily 5's, and with some 4's and 6's; their item scores cluster tightly around 5. Careless responders, instead, may deviate from these purposeful patterns in one of two ways: producing overly repetitive response sets (e.g., all 6's) or producing overly scattered response sets (e.g., 1's and 6's). Scattered response sets may even be sequential (e.g., 1, 2, 3, 4, 5, 6, 5, 4, 3. . .). One common method for flagging overly repetitive response patterns with too little variability is the longstring index (Costa Jr & McCrae, 2008; Huang et al., 2012), quite simply the maximum number of times a single item was endorsed consecutively. Overly scattered responses with too much variability have been assessed using within-person standard deviations (WS; Lang et al. (2019), Marjanovic et al. (2015)). One potential limitation of WS is that it conflates variability in response *type* with variability in response *value*; WS scores will be considerably higher if a person selects 1's and 6's compared to a person who selects 2's and 5's despite both people having identical levels of variability in response type (e.g., both chose only two options). In this paper, I propose the application of an entropy formula for assessment of variability in a response set. As described below, the formula produces a single value that reflects variability in response type only, independent of value, and thus may be a more direct assessment of overly scattered response sets than the WS. Moreover, as discussed below, one of the proposed benefits of the response entropy index is that it simultaneously assesses both too much and too little variability.

Some of the previous studies using multiple indicators of CR have used factor analyses to examine underlying class structures across these metrics. Meade and Craig (2012) found a three-factor solution in which regression type indices (i.e., PS, PA, EO, MD) comprised one factor, four self-report measures of careless responding (i.e., asking participants how attentive they were) comprised the second factor, and two versions of the longstring index comprised the third factor. Consistently, both Grau, Ebbeler, and Banse (2019) and Huang et al. (2012) found the longstring index to load on to a separate factor from regression type indices. I am not aware of studies examining the factor structure of person-fit or variability type indices other than the longstring index,

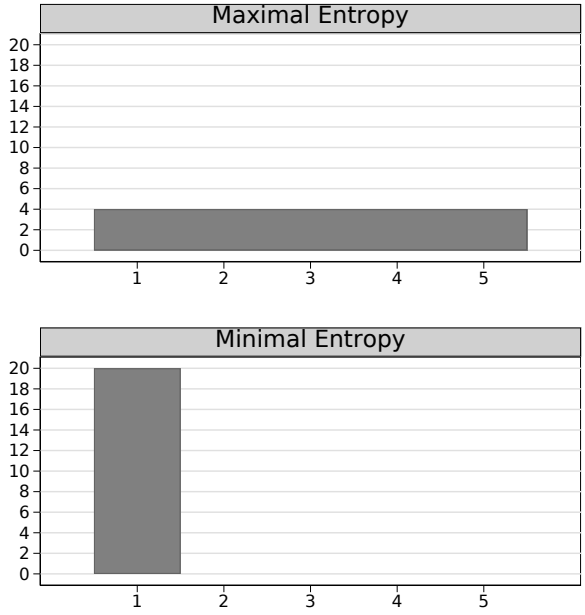

Figure 1. Maximal and minimal entropy distributions

in relation to regression type measures. Overall, the factor analyses that have been done do support the distinction of regression type from the variability type longstring index. I propose the response entropy index as new measure of variability for detecting careless responding.

1.2 Entropy: Application to Careless Responding

Entropy is the idea that over time a system will move towards disorder as a natural state. With time, a ceramic cup will chip, crack, and eventually break down into its raw materials. In measurement, entropy quantifies where a system falls on a continuum of absolute order to absolute disorder. An entropy formula thus provides a single numeric value to represent the system's location on this bipolar continuum and although perhaps most native to physics and thermodynamics (Daprati, Sirigu, Desmurget, Martinelli, & Nico, 2019) this formula has wide application in the sciences. As just a few examples, entropy calculations have been applied to the study of the influence of psychedelics on more disordered (less cognitively constrained) states of consciousness (Carhart-Harris et al., 2014), the level of disorder of button presses on a novel computer task requiring high cognitive engagement (Daprati et al., 2019), and degrees of racial diversity (i.e., disorder) relative to homogeneity (i.e., order) among peer groups (Quillian & Redd, 2009; Tawa, 2017a).

In this application, entropy is used to determine the degree of disorder in a person's response choices on a set of survey questions. A purposeful response set is typically nei-

ther too ordered nor too disordered. Again, consider the example above about a person who is moderately to severely depressed; we would reasonably expect this person to respond on a depression inventory with primarily with 5's, and with some 4's and 6's. A computed entropy score ranges between 0.0 and 1.0, with a maximal entropy score being reached when a person's responses are as diffused across item options as possible. In the case of a 20 item scale, each with five response options, a maximal entropy score would be obtained if a person endorsed four "1's," four "2's," four "3's," four "4's," and four "5's." A minimal entropy score would be obtained if a person endorsed twenty "1's" or any other single number (see Figure 1). Thus, high and low response entropy scores could indicate a person who has respectively either scattered random responses across survey items or constrained responses to a single response type, and in either case, could indicate CR (see section 3.3 for the exact computation of the RE index). While other variability type metrics of CR assess one type of careless responding-overly consistent (i.e., the longstring index) or overly scattered (i.e., within-person standard deviation)-the RE index simultaneously assesses both poles of this continuum. In the current study, I examine the effectiveness of the RE index in relation to other indices of CR for detecting careless responding and the possible factor structures underlying these indices.

2 Study 1

In this study, three types of bogus sets (BS) of data using both computer simulation and human subjects were generated and then sequentially implanted into a true and clean dataset. Indeed, when researchers post their surveys online, humans may not be the only source of CR (**buchanan18**Huang et al., 2012). Programmable survey "bots" can be developed to populate online survey fields with random values to take the appearance of a human completing a survey (**buchanan18**). Thus it is appropriate that many of the above referenced studies examining CR have used computer simulated CR data. Nonetheless, people may also be sources of CR and thus it is also important to empirically examine the effectiveness of these indices for human carelessness. Computers programmed to return random responses are unlikely to generate longstrings or overly consistent patterns; on the other hand, humans are actually poor generators of randomness even when they intend to be (Nickerson, 2002).

First, I hypothesize (H_1) that the RE index will make an independent contribution to the detection of BS from the eight other previously established indices of CR. Second, I hypothesize (H_2) that the RE index will be particularly effective for detecting human generated CR given the capacity of the RE index for detecting overly consistent responding. Third, I hypothesize (H_3) that the total of nine indices examined in this study will comprise three latent factors as conceptualized in the review above: regression type, person-fit type, and variability type. Lastly, I hypothesize (H_4) that each of the three factors will make independent contributions to the prediction of BS.

3 Method

In this section, I describe my approach to developing the clean data, and the computer and human BS data. I then describe the calculation of the nine indices of careless responding.

3.1 Constructing a Clean Dataset

The "clean" dataset was derived from a previously collected social science dataset for a study examining systems beliefs and prejudice. Calculations were based on a single scale within the dataset, the colorblind racial attitudes scale (CoBRAS; Neville, Lilly, Duran, Lee, and Browne (2000)). The CoBRAS is a 20-item, Likert scaled survey, with item endorsements ranging from 1 (strongly disagree) to 6 (strongly agree) and is a good representation of a social science measure in scale length and response options (Worthington & Whittaker, 2006). The survey was completed by 598 adults recruited over the internet on job posting sites (e.g., Craigslist) and through Amazon's MTurk. Before completing the survey, participants read the stern warning: "Your responses may be screened by computer software to detect random or improbable response patterns. If random or improbable response patterns are detected, your survey may be removed from the pool of potential gift certificate winners. This strategy is to eliminate survey response completion by people without actually reading the questions. As long as you read each question and answer each question as honestly as you can, your responses will be accepted!"

Response times were then examined in a further effort to clean these data without employing the post hoc metrics. I was particularly concerned with protocols completed too rapidly (rather than too slowly) as fast protocol completion times have previous been found to reflect careless data (Huang et al., 2012; Wise & Kong, 2005). Eighty-four protocols were found to be completed in under 8.5 minutes-less than half of the median time (17.52 minutes) to complete the protocol-and were removed. In addition, manual examination of response times revealed 14 protocols that took over 100 minutes to complete. Although it is possible that larger response times may result from participants taking a break and returning to the survey, I removed these data given that they were excessively lengthy. The final clean data set is comprised of 500 participants.

Among the 500 participants in the clean dataset, self-identified genders included: 184 male (36.8%), 306 female (61.2%), and 10 missing or other (i.e., transgender; 2.0%). Self-identified racial memberships of participants included: 41 Asian (8.2%), 42 Black (8.4%), 44 Latino (8.8%),

327 White (65.4%), and 46 multiracial or other (i.e., Native American; 9.2%). Forty-five participants (9%) identified as immigrants. Participants' average age was 31.55 (SD = 13.70).

3.2 Construction of Bogus Sets (BS)

I then created three additional datasets, each using a different method of deriving falsified responses: 1) computer-generated uniform random responses; 2) computer-generated normalized random responses; and 3) human-generated careless responses. For the computer-generated uniform random responses, 20 scores for each "participant" (n = 100) between 1 and 6 were drawn each with equal probability. For the computer-generated normalized random responses, 20 scores for each "participant" (n = 100) between 1 and 6 were drawn from a normalized distribution with a specified mean of 3.5 (i.e., the midpoint) and standard deviation of 1.25[1]. A standard deviation of 1.25 would result in 99% of scores occurring with the range of 0.5 (rounded up to 1) to 6.4 (rounded down to 6). The 1% of randomly drawn scores that were less than 0.4 (rounded down to 0) and greater than 6.5 (rounded up to 7) were recoded as 1 and 6 respectively.

The final BS dataset, was created with real participants (n = 100) from Amazon's MTurk. Similar to procedures reported in Leiner (2019), participants were asked to carelessly respond to survey items, in this case, specific to the systems beliefs and prejudice survey with the following instructions: "While taking this survey I would like you to imagine that the primary reason you agreed to participate in this study was to earn a little money or course credit. You are not actually interested in the content of the survey. Perhaps you are even rushed for time and simply want to get through the survey questions as quickly as possible." Among the 100 participants in this BS dataset, self-identified genders included: 71 male (71%), 27 female (27%), and 2 missing or other (i.e., transgender; 2.0%). Self-identified racial memberships of participants included: 22 Asian (22%), 6 Black (6%), 6 Latino (6%), 55 White (55%), and 11 multiracial or other (i.e., Native American; 11%). Nineteen participants (19%) identified as immigrants. Participants' average age was 31.6 (SD = 9.37). Each of the BS datasets was then sequentially embedded into a copy of the clean dataset, comprising a total of three datasets each with 500 clean responses and 100 BS responses[2].

3.3 Calculation of Metrics of Careless Responding

Response Entropy (RE) Index. The RE index is computed by taking the negative of the log of the proportion of each type of response (e.g., proportion of "1's" and "2's") endorsed by the participant, multiplied by the proportion itself, and summing the products. The RE index is calculated for each participant (i) and is represented by the following equation, where K is the number of response types available (in this case, always 6), and $p_k i$ is the proportion of responses named from each response type:

$$\mathrm{RE}_i = -\sum_{k=1}^{K} P_{ki} \log P_{ki}$$

Because a proportion is determined for each value endorsement relative to the total number of a participant's responses (e.g., the proportion of 2's relative to a total of 20 item responses), and because proportions are computed for each possible value (e.g., the proportion of 1's, 2's, 3's, 4's, 5's, and 6's), the RE index will work for measurement scales with any number of response options. Because the RE index is determined independent of the content of items, reverse scored items should be kept in their original form. This will enable detection of both overly scattered and overly consistent responses. For example, if a participant records a "1" for every response in a survey this response set should be characterized as overly consistent; however, reverse scoring items will artificially add variability in response types. High and low RE index scores would result from a person with unusually high or low levels of variability in their item responses and could indicate careless responding.

Within-person Standard Deviation (WS). Within-person standard deviation is operationalized as the standard deviation of a single participant's item responses within a scale (Lang et al., 2019; Marjanovic et al., 2015). Higher standard deviations would likely result from a person who endorses more extreme responses (e.g., 1's and 6's); higher variability could indicate careless responding. Following recommendations by Marjanovic et al. (2015) all reverse coded items are reverse scored before calculating the WS.

Long-String (LS) Index. The LS index is operationalized as the longest string of the same value within a survey. Higher scores on this index could indicate an overly consistent response pattern that characterizes some approaches to careless responding (Costa Jr & McCrae, 2008; Huang et al.,

[1]The mean and standard deviation values were based on maximizing the likelihood of a normal distribution; picking a mean of 3.5 as the midpoint of the 1 to 6 range would allow for equal probability of instances of higher and lower values and less likelihood of a positively or negatively skewed distribution. The determination of the standard deviation of 1.25 was based on the induction that with that range of scores, 99% of the data would fall within the range of 1 to 6. Specifying a higher standard deviation would result in scores falling outside of the range and specifying a lower standard deviation would result in fewer 1's and 6's and a more leptokurtotic distribution.

[2]Across all analyses, with the exception of a multinomial logistic regression, I treated each dataset separately rather than attempting to combine them by creating a multilevel categorical outcome variable. Combining datasets appeared to compromise power given that by combining datasets, 300 of a total 800 observations were falsified scores.

2012). In the case of this analysis it is the longest string of values within the participants' responses to the CoBRAS.

Psychometric Synonym (PS) Index. The PS index is constructed by first examining a correlation matrix of items within a scale and selecting the most strongly *positively* correlated item pairs across a sample (Meade & Craig, 2012). Within-person correlations across item pairs are then computed. Low individual-level *r* values suggest a person is who responds inconsistently to conceptually similar items across a scale and could thus indicate CR. Meade and Craig (2012) recommended basing PS index scores from item pairs that correlate at higher than 0.60. In the current dataset, only six pairs were higher than 0.60, thus the 10 most strongly positively correlated item pairs in the CoBRAS scale were determined for computation of the PS index.

Psychometric Antonym (PA) Index. The PA index is constructed by first examining a correlation matrix of items within a scale and selected the five most strongly *negatively* correlated (Goldberg & Kilkowski, 1985; Meade & Craig, 2012) item pairs across a sample. Within-person correlations across item pairs are then computed. Like the PS index, low individual-level *r* values suggest a person is who responds inconsistently to conceptually similar items across a scale and could thus indicate CR. In the case of the PA index, Meade and Craig (2012) recommended basing PA scores from item pairs that correlate at stronger than −0.60. In the current dataset, only one pair of items were stronger than −0.60, thus the 5 most strongly negatively correlated item pairs in the CoBRAS scale were determined for computation of the PA index.

Even-Odd (EO) Index. The EO index is constructed by constructing item pairs based on all even and odd numbered items within a scale, and in this case, the CoBRAS scale. Within-person correlations across item pairs are then computed. Johnson (2005) recommends that the EO index score be adjusted using the Spearman-Brown split-half prophecy formula which corrects for probability of a lower reliability score with scales with fewer items. This correction, however, can lead to correlation estimates less than −1.0; following Meade and Craig (2012) recommendation I thus recoded scores of less than −1.0 to −1.0. Corrected individual-level *r* values suggest a person is who responds consistently to conceptually similar items across a scale, thus low EO scores may indicate careless responding.

Mahalanobis Distance (MD). MD is a multivariate outlier analysis and is operationalized as the distance of a single participants' data points to the centroid of the overall regression pattern within the sample (Mahalanobis, 1936; Meade & Craig, 2012; Niessen et al., 2016). The centroid is determined as the location in a multivariate space where are all means from all variables included in the analysis intersect. To run MD in SPSS, any other scaled variable is selected and entered as the dependent variable (Meade & Craig, 2012), although the MD is computed based on only the independent variables entered. In this case, I entered all of the CoBRAS items as independent variables and active survey time as the dependent variable (random active time values were generated for BS data). Higher MD scores could indicate CR.

Normed Guttman Errors (G_n^p). A Guttman error occurs when a less probable item is answered affirmatively and a more probable item is answered negatively. In the case of polytomous item scales, Guttman errors are determined for each "item step" (Sijtsma & Molenaar, 2002): the respondent's decision making processes for each possible item answer. For example, consideration of a response of "1" or "2" is the first "item step," consideration of a response of "2" or "3" is the second item step, and so on. A Guttman error occurs each time an item step results in a less probable response being endorsed in favor of a more probable response. Thus, multiple Guttman errors can and do occur within each item. Guttman errors were computed for this sample using the "Profit" package in 'R' (Meijer et al., 2016). For this study, Tendieros (personal communication) recommended I use the normed version of the GE statistic. In the abbreviation for Guttman errors (G_n^p), the superscript "p" indicates it is based off polytomous rather than dichotomous items and the subscript "n" indicates the normed version.

Standardized Log-likelihood (L_z^p). An IRT model determines the probabilities of specific responses to items as a function of a person's trait level, belief endorsement, or degree of knowledge on a subject. In the case of the CoBRAS, a person with a high level of endorsement of color-blind racial ideology would have a predictable response type for each item; for example, imagine the probabilities for each of the response types on item number 1 are: 0.1, 0.1, 0.1, 0.2, 0.3, 0.2; for this item a response of "5" is the most probable. The IRT model determines probabilities for each item. Thus, the standardized log-likelihood is an estimation of the extent to which a specific respondent's pattern of responses fits with the overall response pattern of the sample. L_z^p was also calculated using "Profit" (Meijer et al., 2016). In the abbreviation L_z^p, the superscript "p" indicates it is based off polytomous items and the subscript "z" indicates the standardized version.

4 Results

Preliminary analyses were run with the original true and clean dataset to examine descriptive statistics (see Table 1), demographic variability, and intercorrelations among metrics. A multinomial logistic regression was applied to a dataset constructed of the original clean data and all three types of bogus sets (BS), and was used to estimate the strength of nine different indices for predicting each type of BS. Confirmatory factor analysis was then examined with the original clean data and was used to test the hypothesized latent constructs comprising the nine indices. Lastly, three separate structural regression analyses, each conducted from

Table 1
Descriptive Statistics for Metrics of Careless Responding

	MD[a]	PS[b]	PA[c]	EO[d]	LS[e]	RE[f]	WS[g]	G_n^{p}[h]	L_z^{p}[i]
Minimum	4.25	−0.67	−1.00	−1.00	1.00	0.00	0.00	0.00	−9.40
Maximum	88.00	0.70	0.94	0.95	20.00	0.77	6.58	0.96	2.40
Mean	19.96	0.06	−0.73	−0.01	2.64	0.62	1.82	0.21	0.16
Std. Dev.	12.94	0.31	0.33	0.60	1.94	0.12	1.18	0.14	1.50
Skewness	1.69	0.02	**2.21**	−0.34	**7.00**	−1.87	1.16	1.76	−2.01
Kurtosis	**3.86**	−0.83	**6.08**	−1.14	**57.71**	**5.55**	1.86	4.11	6.75
N	500	443	495	485	500	500	500	500	500

Values in bold represent significant skews.
[a] Mahalanobis Distance [b] Psychometric Synonym Index [c] Psychometric Antonym Index
[d] Even-Odd Index [e] Long String Index [f] Response Entropy Index [g] Within-person Standard Deviation
[h] Normed Guttman Errors [i] Standardized Log-Likelihood.

datasets comprised of the original clean data plus one of the three BS datasets, were then used to examine the independent contribution of latent factors on prediction of BS.

4.1 Preliminary Analyses

Descriptive Statistics and Normality. When there is no variability across a within-person variable (e.g., when the same response is selected for all of the odd items on the EO index) regression type indices (PS, PA, and EO) cannot be computed; thus for these scales there are fewer than 500 observations. Skewness and kurtosis were examined for each variable using the clean dataset. Skewness of less than −2 or greater than +2, and kurtoses of less than −7 or greater than +7 have been found to result in serious biases in confirmatory factor analyses (Curran, West, & Finch, 1996). Thus, I took steps to transform skewed and kurtotic variables.

MD, PA, LS, and G_n^p were all significantly positively skewed and L_z^p was negatively skewed. In order to transform positively skewed variables I followed Tukey's ladder (Tukey, 1977), progressing from logarithmic (log10), to square root, to cube root, and to reciprocal transformations. Prior to transformation of positively skewed variables I added a value of 10 to variables with negative values (i.e., PA). Log transformations of MD and G_n^p normalized and were retained. The reciprocal of both LS and PA normalized. RE approached negative skewness and L_z^p was negatively skewed. Again following Tukey's ladder, for negatively skewed variables I progressed from squaring to cubing. The square of RE and was thus retained; the square and cube of L_z^p exacerbated the skew thus I retained the original variable (which was skewed at −2.01). The final variable forms included the original PS, EO, and L_z^p, the log of MD and G_n^p, the reciprocal of LS and PA, and the square of RE (see Table 2).

Demographic Variability. Next, using the clean dataset, I examined how participant race, immigration status, gender, and age varied in relation to each index of CR. A Multivariate Analysis of Variance (MANOVA) comparing participants' race on the nine indices of CR was significant [Wilks's $\lambda = 0.79; F(27, 1125.04) = 3.85; p < 0.01; h_p^2 = 0.08$]. Follow-up univariate tests[3] revealed that Asian, Black, and Latino/a participants had significantly higher MD scores than Whites; Asian participants had significantly higher PA scores than Whites; Black participants had significantly lower PS scores than Whites; Asian, Black and Latino/a participants had significantly higher WS scores than Whites; and White participants had significantly higher L_z^p scores than Asians, Blacks, and Latino/as. A MANOVA comparing participants' gender on the nine indices of CR was not significant [Wilks's $\lambda = 0.97; F(9, 422) = 1.44; p > 0.05; h_p^2 = 0.03$]. Lastly, a MANOVA comparing participants' immigration status on the nine indices of CR was significant [Wilks's $\lambda = 0.93; F(9, 427) = 3.38; p < 0.01; h_p^2 = 0.07$]. Immigrants had significantly higher MD and WS, and lower EO and L_z^p scores than U.S. born participants (see Table 3 for subgroup means). Age was positively related to EO ($r = 0.13; p < 0.01$) and G_n^p ($r = 0.11; p < 0.01$).

Intercorrelations among Metrics. Correlations among metrics of CR were examined with Pearson's correlations (see Table 4) using the clean dataset.

[3] Follow up univariate ANOVAs were conducted rather than interpreting post hoc multiple comparison from the MANOVA. MANOVAs are not robust for missing data; a missing value on one variable results in listwise exclusion of all other variables. Thus, given the high percentage of values missing as a result of computation of regression type CR indices, reliance on post hoc multiple comparisons would likely result in underestimation of group comparisons on non-regression type indices.

Table 2
Skewness and Kurtosis of Transformed Variables

	MD[a]		PA[c]		LS[e]		RE[f]		G_n^{p}[h]		L_z^{p}[i]	
	SK	KT	SK	KT	SK	KT	SK	KT	SK	KT	SK	KT
Log(10)	**0.11**	**0.22**	2.05	5.17	2.21	9.94	-	-	**−0.08**	**−0.13**	-	-
Square Root	-	-	2.13	5.61	4.50	29.08	-	-	-	-	-	-
Cube Root	-	-	2.10	5.55	3.67	21.07	-	-	-	-	-	-
Recip.	-	-	**−1.91**	**4.36**	**0.85**	**5.41**	-	-	-	-	-	-
Square	-	-	-	-	-	-	**−0.82**	**0.30**	-	-	−3.20	18.13
Cube	-	-	-	-	-	-	-	-	-	-	−12.35	179.40

Final retained values are in bold.
[a] Mahalanobis Distance [c] Psychometric Antonym Index [e] Long String Index [f] Response Entropy Index
[h] Normed Guttman Errors [i] Standardized Log-Likelihood.

4.2 Primary Analyses and Hypothesis Testing

Multinomial Logistic Regression. A multinomial logistic regression was used to test H_1 predicting the independent contribution of RE on detection of BS for each type of CR data. For this analysis only, a single dataset was created combining the three types of BS data with the clean data and specifying data type as a four-level categorical outcome variable. Standardized versions (z-scores) for each variable were used to ease interpretation. With all other indices in the model, the RE index was a significant predictor of human and uniform computerized BS (see Table 5), partially supporting H_1. Negative coefficients for RE suggest that lower scores (resulting from overly consistent responding) predicted human BS, while positive coefficients for RE (resulting from overly scattered responding) predicted uniform computerized BS. H_2 which predicted the particular strength of the RE index for human BS data was not supported; L_z^p appeared to make the strongest contribution with regard to overly consistent responding.

Odds ratios[4] (see Table 5) indicate the probability of categorization of each type of BS for a one-unit increase for each standardized version of the index; for example, the RE index has an odds ratio of 5.93 for prediction of computer uniform data (when other variables are included). This means that for every one unit increase in standardized RE, there is an 5.9 times likelihood of the subject being classified as computerized-normal BS data.

Latent Variable Analyses. Examination of the logistic regressions also suggest that there may be underlying latent constructs operating in the detection of BS responses. For example, PS is a significant stand-alone predictor of BS responses in computer uniform data. However with other variables in the model, PS is no longer a significant predictor. This may indicate that another, latently related construct (e.g., PA) is consuming the variability in BS once captured by PS. I tested the hypothesis (H_3) that the nine indices of CR comprised the three underlying constructs of regression type, person-fit type, and variability type measurements.

For latent variable analyses, first a measurement model was examined with the clean dataset using confirmatory factor analyses[5]. For skewed variables, the transformed variable versions were used for this analysis (see Table 2). Assessment of fit was based on five model-fit indices and their guidelines: The ratio of the Chi-Square statistic to degrees of freedom should range between 1 and 3; Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square (SRMR) should approach statistical significance ($p < 0.05$); and Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) scores should reach between 0.90 and 0.95 (Hu & Bentler, 1999; Jackson, Gillaspy Jr, & Purc-Stephenson, 2009). A three-factor solution based on the theoretical distinctions between the various indices described in the literature review [i.e., regression type (MD, PS, PA, EO), variability type (RE, LS, WS), and person-fit type (G_n^p, L_z^p)] performed better across multiple fit indices than a one-factor solution in which all items were specified as a single factor; however, the fit was still relatively poor (see Table 6). Examination of correlations between items (see Table 4) suggests that MD was highly related to person-fit type indices. Thus, I removed MD from the regression type factor and included in the person-fit factor. This model was improved. Further examination of correlations suggested allowing WS to correlate with person-fit items (i.e., MD, G_n^p, L_z^p), and this model was again improved and reached statistical adequacy. This measurement model was then applied as a predictor of BS in structural regression models.

[4] Sensitivity and specificity analyses were not conducted because they are redundant with information provided in odds ratios (Simel, Easter, & Tomlinson, 2013).

[5] I also attempted to run an exploratory factor analysis using principal axis factoring with all nine items, but it did not converge. Principal axis factoring was selected since at least one item did not normalize. A promax (oblique) rotation was specified as items do correlate.

Table 3
Means and standard deviations for demographic subgroups

		MD[a]		PS[b]		PA[c]		EO[d]		LS[e]		RE[f]		WS[g]		$G_n^{p\text{h}}$		$L_z^{p\text{i}}$	
	n	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Gender																			
male	184	20.88	11.74	**0.02***	0.32	−0.71	0.32	**0.07***	0.61	**2.84***	2.66	0.61	0.13	**1.86**	1.17	**0.23***	0.16	0.02	1.59
female	306	19.57	13.66	**0.09***	0.30	−0.74	0.33	**−0.05***	0.60	**2.48***	0.92	0.63	0.11	**1.81**	1.16	**0.20***	0.13	0.23	1.45
Race																			
Asian	34	**23.55***	11.52	0.07	0.32	**−0.60***	0.38	−0.15	0.70	2.93	1.47	0.61	0.11	2.15*	1.22	0.20	0.11	**−0.33***	1.47
Black	39	**27.03****	15.27	**−0.07***	0.26	−0.74	0.29	−0.03	0.55	2.79	1.20	0.62	0.10	2.41**	1.07	0.19	0.11	**−0.56****	1.55
Latino	36	**25.64****	16.59	0.08	0.29	−0.72	0.35	−0.03	0.62	2.86	2.78	0.58	0.15	**2.31****	1.46	0.23	0.12	**−0.55****	1.78
White	288	**18.05*****	12.05	**0.07***	0.31	**−0.74***	0.32	0.01	0.60	2.53	1.78	0.62	0.12	**1.64*****	1.09	0.21	0.16	**0.42****	1.34
Birthplace																			
US born	455	**19.49****	12.90	0.06	0.30	**−0.74****	0.32	0.01	0.60	2.62	1.96	0.62	0.12	1.77**	1.14	0.21	0.15	**0.22****	1.48
immigrant	45	**24.74****	12.50	0.10	0.31	**−0.60****	0.34	−0.15	0.65	2.93	1.66	0.62	0.12	2.36**	1.40	0.21	0.11	**−0.52****	0.16

Values in bold represent significant skews.
[a] Mahalanobis Distance [b] Psychometric Synonym Index [c] Psychometric Antonym Index [d] Even-Odd Index [e] Long String Index [f] Response Entropy Index
[g] Within-person Standard Deviation [h] Normed Guttman Errors [i] Standardized Log-Likelihood.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 4
Correlations among Metrics of Careless Responding

	1	2	3	4	5	6	7	8	9
1. Mahalanobis Distance	1.00	−0.01	0.38**	−0.12**	0.18**	−0.04	0.86**	0.52**	−0.84**
2. Psychometric Synonym Index	-	1.00	−0.15**	−0.13**	0.11*	0.00	0.16**	0.02	−0.01
3. Psychometric Antonym Index	-	-	1.00	−0.14**	0.17**	−0.05	0.20**	0.10*	−0.33**
4. Even-Odd Index	-	-	-	1.00	−0.07	0.07	−0.08	0.05	0.09*
5. Long String Index	-	-	-	-	1.00	−0.56**	0.26**	0.01	−0.37**
6. Response Entropy Index	-	-	-	-	-	1.00	−0.01*	−0.14**	0.20**
7. Within-person Standard Deviation	-	-	-	-	-	-	1.00	0.33**	−0.80**
8. Normed Guttman Errors	-	-	-	-	-	-	-	1.00	−0.38**
9. Standardized Log-Likelihood	-	-	-	-	-	-	-	-	1.00

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

4.3 Structural Regression Models

Three structural regression models were conducted to examine the impact of the three factor model on the BS outcome. For each regression model, a dataset was developed comprised of the clean data and one type of BS data. Fit statistics for all models were marginal (see Table 7). Regression parameters for each dataset are included below in Table 8. For human BS data, variability and person-fit emerged as significant predictors; for computer-uniform BS, variability and regression emerged as significant predictors; and for the computer-normalized BS data variability and regression indices emerged as significant predictors. Variability was the only class of indices that detected all three types of BS data.

5 Discussion

In the first study, the response entropy (RE) index was compared to other measures of detecting careless responding by assessing each measures' ability to identify false data created by the researcher and implanted into true datasets. There are some reasonable challenges to this method, for example, despite efforts to create a true and clean dataset, it is not only possible but also *likely* that careless responders still exist in the true and clean dataset. An alternative approach taken by careless responses researchers (e.g. Meade & Craig, 2012) is to simulate clean data, for example, by imputing response sets that resemble responses for carefully attending participants. While the strength of the simulation approach is that the comparative data is more purely clean, there is also an obvious criticism; simulated "clean" responses do not have the type of unique variability in motivation, attention, and spuriousness that human beings have. In fact, I would argue that because of the human variability existing within true comparative data, the approach used in this study is a more stringent test of the effectiveness of careless response indices than approaches that use simulated clean comparative data.

Based on this study, the response entropy (RE) index appears to be a viable, but not exclusive, alternative to previously used post hoc indices of detecting careless responding. However, as a comparatively user-friendly metric to implement with capacity to simultaneously assess overly consistent and scattered responses, the RE index may be a particularly attractive option for social scientists using survey methods. As a brief review, preliminary analyses determined that the RE index was the only metric that was not significantly related to any of the demographic variables assessed. Multinomial logistic regression analyses determined that the RE index significantly predicted human careless and computerized-uniform BS data, even with all other metrics included in the model. The multinomial logistic regression also suggested that there may be underlying factors across metrics of CR; for example, although WS detected computerized-uniform BS data alone (i.e., step 0), when entered with all other variables in the model (i.e., step 1) the effect disappeared. Hunsley and Meyer (2003) point to the need for aggregation approaches in incremental validity studies precisely because it is common for some underlying construct to account for associations between multiple predictors and the outcome variable examined. For example, I suspected that the WS contribution to the detection of BS disappeared in step 1 because it is subsumed within the construct of variability-based metrics, along with LS and RE. Indeed, a confirmatory factor analysis determined an adequate fit for the hypothesized latent construct model (if MD was removed from the regression type factor into the person-fit factor). Moreover, structural regression models suggested that the three latent constructs of variability, regression, and person-fit each made unique contributions to the detection of BS data.

Nonetheless, the use of factor analyses in the current study may raise some reasonable challenges, for example, there is data dependency across the CR measures. In fact, this challenge could be raised not only for the current study, but its predecessors that have also employed factor analyses in attempt to understand latent constructs within careless response measures (Grau et al., 2019; Huang et al., 2012;

Table 5
Multinomial Logistic Regression

	Human				Uniform				Normal			
	Coef.		OR		Coef.		OR		Coef		OR	
	Step 0	Step 1	Step 0	Step 1	Step 0	Step 1	Step 0	Step 1	Step 0	Step 1	Step 0	Step 1
Mahalanobis Distance												
b	0.44**	−1.60**	1.55	0.21	1.48**	0.27	4.41	1.30	0.87**	0.79	2.38	2.20
S.E.	0.13	0.34	-	-	0.14	0.37	-	-	0.12	0.37	-	-
Psychometric Synonym Index												
b	−0.06	−0.43**	0.94	0.65	−0.59**	−0.36	0.55	0.70	−0.61**	−0.51**	0.54	0.60
S.E.	0.12	0.16	-	-	0.12	0.17	-	-	0.12	0.15	-	-
Psychometric Antonym Index												
b	1.22**	0.95**	3.38	2.59	1.78**	1.50**	5.93	4.42	1.75**	1.50**	5.77	4.42
S.E.	0.14	0.17	-	-	0.15	0.18	-	-	0.15	0.16	-	-
Even-Odd Index												
b	−0.41**	−0.20	0.67	0.82	−0.01	0.22	1.00	1.24	−0.03	0.21	0.97	1.23
S.E.	0.11	0.15	-	-	0.11	0.16	-	-	0.11	0.15	-	-
Long String Index												
b	0.23**	−0.98**	1.26	0.38	−0.24	0.02	0.79	1.02	−0.07	−0.36	0.93	0.70
S.E.	0.08	0.34	-	-	0.22	0.41	-	-	0.15	0.35	-	-
Response Entropy Index												
b	−0.60**	−0.66**	0.55	0.52	1.94**	1.78**	6.93	5.93	0.26	0.27	1.29	1.31
S.E.	0.10	0.22	-	-	0.27	0.37	-	-	0.15	0.26	-	-
Within-person Std. Dev.												
b	−0.26**	−0.07	1.30	0.93	0.31**	−0.14	1.37	0.87	−0.38**	−1.2**	0.68	0.31
S.E.	0.10	0.18	-	-	0.10	0.22	-	-	0.14	0.26	-	-
Normed Guttman Errors												
b	0.17	0.38	1.18	1.46	0.48**	0.02	1.62	1.02	0.08	−0.68	1.09	0.51
S.E.	0.11	0.21	-	-	0.10	0.26	-	-	0.12	0.30	-	-
Standardized Log-Likelihood												
b	−1.20**	−2.30**	0.30	0.10	−1.40**	−1.20**	0.25	0.30	−0.85**	−0.65	0.43	0.52
S.E.	0.13	0.32	-	-	0.13	0.34	-	-	0.13	0.34	-	-

AIC: 1038.58
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table 6
Model Fit Statistics for Measurement Model

Model	X^2	$\frac{X^2}{df}$	CFI	TLI	RMSEA	SRMR
Model A	288.62^{**}	10.69	0.83	0.77	0.15	0.08
Model B	260.65^{**}	10.86	0.85	0.77	0.15	0.09
Model C	207.75^{**}	8.66	0.88	0.82	0.13	0.09
Model C (mod)	95.62^{**}	4.55	0.95	0.92	0.09	0.08

$^{*}\ p < 0.05$ $^{**}\ p < 0.01$ $^{***}\ p < 0.001$

Table 7
Model Fit Statistics for Structural Regressions

Model	X^2	$\frac{X^2}{df}$	CFI	TLI	RMSEA	SRMR
Human	231.77^{**}	8.54	0.87	0.79	0.12	0.08
Comp. Uni	125.10^{**}	4.63	0.96	0.94	0.08	0.06
Comp. Norm	118.79^{**}	4.40	0.96	0.93	0.08	0.06

$^{*}\ p < 0.05$ $^{**}\ p < 0.01$ $^{***}\ p < 0.001$

Huang et al., 2015; Meade & Craig, 2012). Perhaps future studies can seek alternative methods of aggregation in order to heed Hunsley and Meyer (2003) warning while simultaneously addressing data dependency. In addition, the move of MD from regression type indices to person-fit indices was in this case, purely empirically derived, and a more conceptual understanding of this fit should be explored. Future researchers should examine how this factor structure holds up with their own dataset which may vary, for example, on distribution, item numbers, and value ranges.

Despite these reasonable challenges, within the current sample, structural regression models suggested that the three latent constructs of variability, regression, and person-fit each made unique contributions to the detection of BS data. Variability type indices were the only class of metrics to predict all three types of BS data. In what follows I examine the capacity of each class of metric-with a focus on the contribution of the RE index-towards detection of each type of BS data.

5.1 Human Careless Responses

Human careless responders are more likely to overly consistently endorse a single response than computer programs that are designed to populate surveys with random numbers (Nickerson, 2002). For example, a person may rush through a survey by recording values of "3" for all items. Thus, I hypothesized that the RE index would be particularly strong in predicting human BS; this was supported in both the structural regression and multinomial logistic regression models. In the structural regression model both variability and person-fit indices emerged as unique predictors for detection of human careless BS data. Among the three available variability type metrics (i.e., RE, LS, and WS), the multinomial logistic regression points particularly to the LS and RE index as significant predictors of human careless BS. Consistent with my hypothesis, the direction of the negative z score[6] for the RE index suggests that in this instance, this metric is contributing to the detection of overly consistent responding (i.e., low variability). Person-fit indices (specifically MD and L_z^p) then, are likely contributing to the detection of other types of aberrant responses that do not fit with the more modal response patterns in the dataset. L_z^p was notably the strongest predictor of human careless responses in the multinomial regression model. The negative z score in the multinomial regression suggests that data categorized as BS results in lower L_z^p scores than non-BS data; BS data are less likely to "match" the overall high probability responses for each item in the dataset. Regression type indices can also certainly be effective, as both PA and PS emerged as significant predictors of human careless BS in the multinomial regression; however as a class of indices they do not appear to make a contribution to the detection of human careless BS that is unique from the contributions of variability and person-fit. Evaluation of the structural regression model for human BS should be interpreted with some caution as the fit of this model was marginal.

[6]Directionality for z scores in the structural regression are not interpreted because some classes of metrics contain metrics in which directionality would be interpreted in the opposite direction. For example, within the variability class, high RE scores would indicate more scatter (i.e., more variability in response types) while high LS scores would indicate less scatter (i.e., less variability in response types).

Table 8
Structural Regression Parameters for BS Outcomes

	Coef.	S.E.	z	Std. (latent)	Std. (all)
Human					
Variability	−0.25	0.05	−4.73**	−0.25	−0.69
Regression	−0.05	0.05	−1.00	−0.05	−0.14
Person-Fit	0.13	0.04	3.28**	0.13	0.35
Uniform					
Variability	−0.09	0.03	−2.99**	−0.09	−0.22
Regression	0.36	0.06	5.77**	0.36	0.92
Person-Fit	−0.05	0.06	−0.84	−0.05	−0.12
Comp. Norm					
Variability	0.07	0.04	1.99*	0.07	0.18
Regression	0.37	0.06	6.59**	0.37	0.95
Person-Fit	−0.05	0.04	−1.14	−0.05	−0.13

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

5.2 Computer-Uniform Responses

Computerized random number generation will approach a platykurtic distribution, thus would likely result in a scattered type response set. As a case study, a set of responses for 10 items ranked on a Likert scale of 1 to 6, that contained 2 1's, 2 2's, 2 3's, 2 4's, 2 5's, and 2 6's, would resemble a completely flat distribution and would yield the highest possible RE score. Thus, it is not surprising that the RE index is particularly adept at detecting overly scattered responses from generated computer-uniform data and emerged as the strongest of all nine CR indices in the multinomial regression. In this case, the positive direction of the z score suggests the RE index is detecting overly scattered response sets. The change in direction of prediction with human careless compared to computer-uniform BS data demonstrates the flexibility of the RE index for measuring both overly consistent and overly scattered responses. The regression type metrics class was also a significant predictor of BS within the structural regression model, and the multinomial logistic regression points particularly to PA as the strongest of the regression type indices towards prediction of BS. It is worth noting that L_z^p was the only other metric to predict computerized-uniform BS in the multinomial regression, although as a class of indices person-fit did not make a contribution to the detection of BS that is unique from the contributions of variability and regression.

5.3 Computer-Normalized Responses

Computer-normal data will approach a normal distribution with mid-range response types occurring more frequently than more extreme responses. Again as a case study, a set of responses for 10 items ranked on a Likert scale of 1 to 6, will likely result in a disproportionate number of 3's and 4's, fewer 2's and 5's, and the least amount of 1's and 6's. Thus computer-normal BS is most likely to produce a response set that resembles a participant who moderately agrees with the content of items. I am not aware of any assessments related to how prevalent this type of falsified data is, I only know that it is conceivable to generate. And, it appears that this type of BS data would be the most difficult for the RE index to detect given that proportions of response types would be most likely to resemble carefully attended human data. Nonetheless, the structural regression model found that the variability type factor is a significant predictor of computer-normalized BS. In the case of computer-normalized data, regression based metrics may be the strongest predictors of BS data. While the RE index and variability class measures evaluate data quality via examination of proportions of response types, regression based measures evaluate data quality based on common patterns of relationships between items within a dataset. For example, in a survey comprised of 20 items, item 4 and item 8 may have a particularly high correlation among the sample. Randomly generated computer data will not be as likely to result in similar within item correlations, even if the distribution of responses resembles that of a carefully attended human response.

In application, a researcher will not know what type of falsified data is contained in their dataset. Thus, to cover all bases, I would recommend that researchers interested in flagging BS responses select at least one variability type measure in order to evaluate *proportionally* aberrant responses and also at least one regression type measure in order to evaluate *relationally* aberrant responses, the latter of which would be particularly valuable for detection of BS with computer-normalized data. To be very thorough, researchers may also wish to include a person-fit type measure which would help detect BS on the basis of aberrant responding due to endorse-

ment patterns of items that are less modal within the sample. In terms of specific metrics within each class of BS, I would recommend: 1) the RE index which emerged as the strongest predictor in its class for detecting BS with human careless data and computer-uniform data; 2) the PA index which emerged in the multinomial regression as a significant predictor of all three types of BS; and 3) the L_z^p which emerged as the strongest predictor in its class for detecting BS with human careless data and computer-uniform data.

Lastly, preliminary findings indicate that with the exception of RE, indices of CR were frequently related to demographic differences. One possibility is that these indices are being confounded with culturally different response styles. If this is the case, it is problematic. The removal of cultural variability in datasets due to confounding cultural values with CR could perpetuate the marginalization of underrepresented demographic groups in social science research. This exploratory finding in study 1, prompted the development of study 2 in which I hypothesized that demographic differences on indices of CR are a result of conflation of CR with cultural values.

6 Study 2

The goal of the second study was to further examine the relations of demographic variability and indices of careless responding (CR) and specifically to further evaluate the hypothesis raised in study 1: that the response entropy (RE) index was less likely than other indices of careless responding (CR) to be conflated with cultural values, and more specifically, culturally diverse responding styles. Previous research examining indices for detecting CR have rarely examined how indices of CR relate to demographic variability. Some of these studies have used simulated data (e.g. Huang et al., 2015; Meade & Craig, 2012) thus analyses of demographic variability is not possible. However, a number of studies have collected demographic data such as participants' gender (Huang et al., 2012; Huang et al., 2015; Johnson, 2005; Marjanovic et al., 2015; Niessen et al., 2016) and participants' race (Bowling et al., 2016; Huang et al., 2015) but did not examine effects. When researchers have analyzed demographic variability on CR, findings have tended to emerge, although not always consistently. Both Grau et al. (2019) and Bowling et al. (2016) found that people with higher levels of education tend to have lower levels of CR. Grau et al. (2019) did not find age or gender to be related to CR, however Schneider et al. (2017) found that younger adults and males (relative to females) were more like to be careless responders.

The focus of the current study is on ethnic cultural demographic variability[7] in CR, of which previous examination has been woefully inadequate. In the first study reported in this article, I found White participants to have lower Mahalanobis distance (MD) and higher standardized log-likelihood (L_z^p) scores than Asian, Black, and Latino/a participants. Grau et al. (2019) offer perhaps the most comprehensive analysis of the relation of culture to CR to date based on a sample of more than 8,000 participants representing 34 different countries. These researchers used an aggregate measure of CR comprised from regression type indices [e.g., psychometric synonym (PS) and antonym (PA) indices], MD, and the longstring (LS) index. CR was then averaged across all participants within each country to enable cross-cultural comparisons; as some examples, US based participants had considerably lower levels of CR ($M = -0.32$) than participants from Ecuador ($M = 0.54$), Guatemala ($M = 0.38$), and Pakistan ($M = 0.29$), however significance testing was not reported.

There are a number of possible explanations as to why demographic groups may differ in CR. For example, in the case of findings of greater CR among those with less education (Bowling et al., 2016; Grau et al., 2019), one likely explanation is that people's lack of familiarity with surveys may result in people misunderstanding or losing interest in surveys. In the case of ethnic cultural variability, it is less clear. One possibility is that diverse cultural value systems influence how people tend to respond to surveys (Bachman & O'Malley, 1984; He, de Vijver, Espinosa, & Mui, 2014; He & Van De Vijver, 2015; Hui & Triandis, 1989; Smith, 2004). For example, agreeableness is highly valued within many Asian cultures as it serves to promote collectivism and harmony (Sue & Sue, 2012), and consequently Asians have been found to provide fewer extreme responses and more "yes-saying" or acquiescent response patterns (Smith, 2004). Both Black (Bachman & O'Malley, 1984) and Latino/a (Hui & Triandis, 1989) participants have been found to have a tendency towards more extreme responding and less midpoint responding than Whites. In turn, these culturally meaningful response styles may impact the extent to which a pattern of responses are deemed careless. To be very clear, this hypothesis raises the possibility that there are actually *not* cultural differences in CR, but rather that culturally diverse response styles are being miscategorized as CR by common indices of CR.

There is some indication that response styles do relate to careless responding, supporting the "cultural confound" hypothesis stated above. In Grau et al. (2019) study, using their aggregate measure of CR, the researchers found that CR was

[7] In this study I use participants' self-reported *race* (i.e., Asian, Black, White) as a proxy for ethnic/cultural groups. While races are socially constructed aggregate groupings comprised of multiple cultural and ethnic groups, ethnic minority psychologists recognize that there are often enough overlaps in cultural values among ethnic groups (e.g., Thai, Japanese, Korean) within a racial group that differences found between racial groups may reflect differences in cultural values. It is thus common within ethnic minority psychology to speak of, for example, "Asian cultural values" (B. S. Kim, Atkinson, & Yang, 1999; Sue & Sue, 2012).

positively correlated with extreme, midpoint and acquiescent responding, and negatively correlated to social desirability. These four response styles (or in the case of social desirability, an influence on response style) loaded onto their own factor from CR, thus while related, they are distinct constructs. Huang et al. (2015), using a similar aggregate measure of CR also found CR to positively relate to midpoint responding.

In this study, I test the "cultural confound" hypothesis using a dataset comprised of survey responses by Asian (n = 59), Black (n = 67), and White (n = 158) participants recruited from a student subject pool and online[8]. Indices of CR were based on responses to the Beliefs about Race Scale (BARS; Tawa (2017b)). Following Grau et al. (2019), I examine CR indices in relation to midpoint, extreme, acquiescent, and social desirability response styles. Specifically, I predict that differences between Asian and White participants' scores on CR indices will be mediated by acquiescent style responding, and that differences between Black and White participants' scores on CR indices will be mediated by extreme style responding.

7 Method

7.1 Participants

Among the 302 participants, self-identified genders included: 111 male (36.8%), 186 female (61.8%), and 5 missing or other (i.e., transgender; 1.7%). Self-identified monoracially identified participants included: 59 Asian (19.5%), 67 Black (22.2%), 158 White (52.3%), and 18 missing or other (9.2%). Forty-one participants (13.6%) identified as immigrants. Participants' average age was 34.5 (SD = 11.04).

7.2 Measures

All indices of careless responding were calculated in the same way as study 1; in this study calculations were determined from the BARS (Tawa, 2017b) a 16 item measure with items on a 1 - 6 Likert scale. Calculation of three of the response styles (midpoint, extreme, and acquiescent) followed procedures in Grau et al. (2019) and were also determined from the BARS, while social desirability was measured using a separate scale (Crowne & Marlowe, 1960).

Midpoint Responding (MDPT). Midpoint responding was operationalized as the frequency of item responses of "3" and "4" on the 1 to 6 Likert scale range (Grau et al., 2019). Greater numbers of "3's" and "4's" would result from a person who tends to use mid-range options.

Extreme Responding (EXT). Extreme responding was operationalized as the frequency of item responses of "1" and "6" on the 1 to 6 Likert scale range (Grau et al., 2019). Greater numbers of "1's" and "6's" would result from a person who tends to use high and low range options.

Acquiescent Responding (ACQU). Acquiescent responding was operationalized as the mean score on all items without reverse scoring reverse worded items (Grau et al., 2019). A higher value should indicate a person who tends to agree with item content even if items are contradictory in content (i.e., reverse and non-reverse worded).

7.3 Social Desirability (SDS)

Social desirability was measured using Crowne and Marlowe (1960) Social Desirability Scale that assesses the extent to which participants tend to misrepresent themselves as a way to manage their self-perception. This scale includes 33 true-false items. A sample item is: "My table manners at home are as good as when I eat out in a restaurant." Higher numbers of "true" responses indicate greater social desirability. In this study, an internal reliability estimate using Kuder-Richardson's formula 20 was minimally acceptable (KR = 0.61) and findings related to this construct should be interpreted cautiously.

8 Results

Preliminary analyses were first run to examine descriptive statistics for all nine indices of CR and four response styles. A factor analysis was run to examine conceptual distinctions between CR indicators and response styles. Analyses of variance examined how participant race, as a proxy for culture, was related to each of the CR indices. When there were effects of participant race, parallel mediation analyses were conducted to examine which of the response styles accounted for the association between participant race and CR.

8.1 Preliminary Analyses

Descriptive Statistics and Normality. Again, when there is no variability across a within-person variable (e.g., all of the odd items for the EO index) regression type indices (PS, PA, and EO) cannot be computed; thus for these scales there are fewer than 302 observations. Among all CR (see Table 9) and response style (see Table 10) variables, only one variable (LS) exceeded a skewness of greater than +2 and kurtosis of greater than +7 (Curran et al., 1996), but this variable could not be improved by transformation so it was retained in its original form.

Intercorrelations among Metrics. Correlations among CR and response style indices were examined with Pearson's correlations (see Table 11). The RE index was positively related to MD, PS, EO, and WS, and negatively related to PA and LS. The RE index was not related to any response styles. With the exception of the LS index, all other CR indices were correlated with at least two response styles (see Table 11).

[8]Regrettably, recruitment source was not tracked in this study thus differences in CR by recruitment source could not be tested.

Table 9
Descriptive Statistics for Metrics of Careless Responding

	MD[a]	PS[b]	PA[c]	EO[d]	LS[e]	RE[f]	WS[g]	G_n^{p}[h]	L_z^{p}[i]
Minimum	2.00	−0.67	−1.00	1.00	1.00	0.00	0.12	−1.70	5.40
Maximum	56.50	1.00	1.00	1.00	16.00	0.76	6.67	−0.10	2.90
Mean	15.92	0.55	−0.23	0.44	4.04	0.53	2.30	−0.85	0.19
Std. Dev.	10.36	0.38	0.65	0.58	2.28	0.15	1.37	0.30	1.45
Skewness	1.46	−0.95	0.53	−1.43	**2.88**	−1.10	0.90	−0.16	−0.97
Kurtosis	2.10	0.14	−1.14	0.97	**11.51**	1.32	0.32	−0.25	1.02
N	302	288	202	294	302	302	302	300	302

Values in bold represent significant skews.
[a] Mahalanobis Distance [b] Psychometric Synonym Index
[c] Psychometric Antonym Index [d] Even-Odd Index [e] Long String Index
[f] Response Entropy Index [g] Within-person Standard Deviation
[h] Normed Guttman Errors [i] Standardized Log-Likelihood.

Table 10
Descriptive Statistics for Response Styles

	MDP[a]	AQ[b]	EXT[c]	SDS[d]
Minimum	0.00	1.00	0.00	3.00
Maximum	16.00	6.00	16.00	33.00
Mean	5.90	3.42	4.76	19.87
Std. Dev.	3.97	0.78	4.62	4.18
Skewness	0.59	0.14	0.84	0.41
Kurtosis	−0.11	0.50	−0.31	1.99
N	302	302	302	302

[a] Midpoint Responding
[b] Acquiescent Style Responding
[c] Extreme Score Responding
[d] Social Desirability Scale

Factor Analysis. A factor analysis was conducted with this sample primarily to confirm that, as in the previous sample, the RE index appeared to be measuring a different latent construct than the regression and person-fit type measures. Additionally, I was interested in determining if the CR indicators were indeed distinct conceptually from the response styles. Exploratory factor analysis was conducted using a maximum likelihood estimation (given that variables were generally normally distributed) with a Direct Oblimin rotation. A correlation between at least one factor pair was higher than the 0.32 threshold recommended by Tabachnick, Fidell, and Ullman (2007) suggesting an oblique factor structure. A Kaiser-Meyer-Olkin score of 0.714 indicated that the sample size was minimally acceptable. Based on Kaiser criterion (eigenvalues of greater than 1.0) and examination of a scree plot, a four-factor model was recommended (Costello & Osborne, 2005). Examination of the factor loadings of the four factors suggest that again RE and LS comprised a single factor; however, this time WS mapped more strongly on to a factor including EXT and MDPT. The person-fit type indices again combined with MD appeared to form a third factor, and a final fourth factor comprised the regression type indices (i.e., PS, PA, and EO) in addition to ACQU and SDS. These findings support that as in study 1, response entropy appears to be measuring a different construct than other measures of careless responding and response styles[9]. Response styles,

[9]Given that the primary purpose of the EFA was to establish the relative independence of the response entropy measures from other careless response indices and response style indices, no further exploration of the factor structure of this dataset was conducted. A CFA determined that the four-factor model recommended by the EFA: factor 1 [(MDPT, EXT, STDEV); factor 2 (RE, LS), factor 3 (G_n^p, L_z^p, MD), and factor 4 (PA, PS, EO, SDS, ACQU)] was a marginally sound fit ($X^2/df = 4.76; CFI = 0.84; TLI = 0.79; RMSEA = 0.14; SRMR = 0.08$), and it was considerably improved relative to a one-factor model ($X^2/df = 13.42; CFI = 0.45; TLI = 0.33; RMSEA = 0.25; SRMR = 0.19$). Again, given that the primary goal of establishing the relative independence of the response entropy variables, I did not pursue further modification for

Table 11
Correlations among Metrics of Careless Responding and Response Styles

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Mahalanobis Distance	1.00	−0.23**	0.08	−0.17**	−0.17**	0.17**	0.60**	0.74**	−0.78**	−0.44**	0.06	0.50**	−0.09
2. Psychometric Synonym Index	-	1.00	−0.17*	0.63**	0.20**	0.12*	0.37**	−0.06	0.18**	−0.39**	−0.26**	0.35**	−0.12*
3. Psychometric Antonym Index	-	-	1.00	−0.32**	0.20**	−0.19**	−0.10	−0.01	−0.03	0.11	0.18*	−0.14*	0.19**
4. Even-Odd Index	-	-	-	1.00	0.10	0.17**	0.30**	−0.11	0.13*	−0.29**	−0.34**	0.25**	−0.17**
5. Long String Index	-	-	-	-	1.00	−0.62**	0.10	−0.12*	0.08	0.03	0.00	0.06	0.08
6. Response Entropy Index	-	-	-	-	-	1.00	0.18**	0.01	−0.05	−0.05	−0.03	−0.09	−0.11
7. Within-person Standard Deviation	-	-	-	-	-	-	1.00	0.55**	−0.48**	−0.67**	−0.13**	0.77**	−0.12*
8. Normed Guttman Errors	-	-	-	-	-	-	-	1.00	−0.75**	−0.51**	−0.04	0.67**	−0.04
9. Standardized Log-Likelihood.	-	-	-	-	-	-	-	-	1.00	0.32**	0.16**	−0.42**	0.07
10. Midpoint Responding	-	-	-	-	-	-	-	-	-	1.00	0.11	−0.67**	0.09
11. Acquiescent Style Responding	-	-	-	-	-	-	-	-	-	-	1.00	0.27**	0.16**
12. Extreme Score Responding	-	-	-	-	-	-	-	-	-	-	-	1.00	0.14**
13. Social Desirability Scale	-	-	-	-	-	-	-	-	-	-	-	-	1.00

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 12
Factor Loadings

	1	2	3	4
Normed Guttman Errors	**0.635**	0.074	0.441	0.134
Standardized Log-Likelihood	**−1.02**	0.033	0.001	0.175
Mahalanobis Distance	**0.519**	0.318	0.448	0.313
Response Entropy Index	−0.075	**0.911**	0.103	−0.391
Long String Index	−0.030	**−0.557**	0.049	−0.126
Within-person Standard Deviation	0.108	0.079	**0.848**	−0.120
Midpoint Responding	0.045	0.029	**−0.771**	−0.036
Extreme Score Responding	0.084	−0.029	**0.818**	−0.097
Psychometric Synonym Index	−0.230	−0.175	0.284	**−0.661**
Psychometric Antonym Index	0.068	−0.070	−0.034	**0.277**
Even-Odd Index	−0.161	−0.163	0.226	**−0.611**
Acquiescent Style Responding	−0.129	0.032	0.057	**0.542**
Social Desirability	−0.045	−0.048	0.010	**0.291**

however, were not always distinct from CR; WS comprised a factor with midpoint and extreme responding, and ACQU and SDS comprised a factor with the regression type indices (see Table 12). Thus, in the mediation analyses below, findings in which the independent variable and mediator variable are members of the same factor should be interpreted cautiously.

8.2 Primary Analyses and Hypothesis Testing

Multivariate Analyses of Variance (MANOVA). A MANOVA comparing participants' race[10] on the nine indices of careless responding was significant [Wilks's $\lambda = 0.80; F(18, 356) = 2.28; p < 0.01; h_p^2 = 0.10$]. Follow-up univariate tests revealed that White and Asian participants' had significantly lower MD scores compared to Blacks; Asian participants had significantly lower PS and EO compared to Whites; Asian participants had significantly higher PA scores compared to Blacks and Whites; and Black participants had significantly higher WS and G_n^p and lower L_z^p scores than Asians and Whites. LS and RE were the only indices unrelated to race (see Table 13).

A Multivariate Analysis of Variance (MANOVA) comparing participants' race on the four indices of response styles was significant [Wilks's $\lambda = 0.94; F(8, 556) = 2.25; p < 0.05; h_p^2 = 0.03$]. Follow-up univariate tests revealed that Asian participants' had significantly higher MDPT scores than Blacks, and significantly higher SDS scores than Whites (see Table 14).

Mediation Analyses. Parallel mediation models were used to examine the hypotheses that cultural differences in careless responding would be better accounted for by diverse response styles. Mediation analyses were examined using the bootstrapping method aided by PROCESS version 3, model number 4 (Hayes, 2013). Bootstrapping provides an estimate of both the direct path (i.e., the relation between the predictor and outcome variable while controlling for the effect of the mediation variables) and the indirect paths (i.e., the path from the predictor to the outcome, through each mediation variable). Each analysis was based on 10,000 resamples of the dataset with a bias corrected 95% confidence interval. In this method, the indirect effect is considered significant at $p < 0.05$ if the provided confidence interval does not contain the value of 0 (Hayes, 2013; Preacher & Hayes, 2008). Mediation models were run for CR indices in which participant race effects were found. Since the RE index is the focus of this paper, this variable was also included even though no direct effect of race was found. Because mediation cannot be conducted with a categorical variable with more than two levels, each model was run twice, once to compare Asian participants to White participants, and once to compare Black participants to White participants (see Tables 15 and 16). Participant race (Asian vs. White or Black vs. White) was the independent variable, all four response styles (MDPT, EXT, ACQU, SDS) were run as mediator variables in parallel, and the eight CR indices were the dependent variables.

Asian participants (coded as 0) had significantly higher

improvement of fit statistics.

[10]MANOVAs were also run to examine the influence of gender on careless response indices [Wilks's $\lambda = 0.95; F(9, 186) = 1.03; p = 0.42; h_p^2 = 0.05$] and response styles [Wilks's $\lambda = 0.99; F(4, 292) = 0.75; p = 0.56; h_p^2 = 0.01$] and immigration status on careless response indices [Wilks's $\lambda = 0.95; F(9, 191) = 1.04; p = 0.41; h_p^2 = 0.05$] and response styles [Wilks's $\lambda = 0.98; F(4, 297) = 1.57; p = 0.18; h_p^2 = 0.02$]; none of these analyses were significant. Age was also unrelated to all careless response and response style indices. Given that these analyses were not the focus of this study, these statistics are not included in the body of the text.

Table 13
Mean differences on CR measures by participant race.

		MD[a]		PS[b]		PA[c]		EO[d]		LS[e]		RE[f]		WS[g]		G_n^{Ph}		L_z^{pi}	
	n	D	S.E.	D	S.E.	D	S.E.	D	S.E.	D	S.E.	D	S.E.	D	S.E.	D	S.E.	D	S.E.
Asian	59*	15.93**	10.83	0.45**	0.42	0.03**	0.65	0.25**	0.67	4.09	2.60	0.53	0.16	1.99**	1.30	−0.91**	0.34	0.28**	1.66
Black	67*	20.45**	12.09	0.51	0.41	−0.32**	0.65	0.41	0.57	3.84	1.78	0.54	0.14	2.57**	1.54	−0.73**	0.26	−0.44**	1.55
White	158*	14.09**	8.76	0.61**	0.34	−0.30**	0.63	0.52**	0.54	4.10	2.42	0.52	0.15	2.06**	1.30	−0.86**	0.28	0.41**	1.27

[a] Mahalanobis Distance [b] Psychometric Synonym Index [c] Psychometric Antonym Index [d] Even-Odd Index [e] Long String Index
[f] Response Entropy Index [g] Within-person Standard Deviation [h] Normed Guttman Errors [i] Standardized Log-Likelihood.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 14
Mean differences on CR measures by participant race.

		MDPT[a]		ACQU[b]		EXT[c]		SDS[d]	
	n	D	S.E.	D	S.E.	D	S.E.	D	S.E.
Asian	59	6.66**	3.67	3.65	0.78	3.86	4.36	21.12**	4.37
Black	67	4.91**	3.73	3.41	0.68	5.30	4.72	20.15	3.84
White	158	6.02	4.14	3.38	0.81	4.86	4.67	19.39**	4.16

[a] Midpoint Responding [b] Acquiescent Style Responding
[c] Extreme Score Responding [d] Social Desirability Scale
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

levels of MD, marginally higher levels of PA, and marginally lower levels of EO and PS than White participants (coded as 1). The relationships between participant race (Asian vs. White) and MD, PS, and EO, were mediated by acquiescent responding such that Asian participants had higher levels of acquiescent responding than Whites, which in turn was negatively related to MD, and positively related to PS and EO. Each of these indirect paths was significant. Acquiescent and socially desirable responding also served as an indirect path between participant race (Asian vs. White) and G_n^p , although the direct effect was not significant (see Table 15). Black participants (coded as 0) had higher levels of MD, G_n^p, and WS and lower levels of PS, EO, and L_z^p than White participants (coded as 1) and these direct effects were significant. The relationships between participant race (Black vs. White) and MD, EO, and WS were also mediated by midpoint responding such that Black participants had lower levels of midpoint responding than Whites, which in turn was negatively related to MD, WS, and EO. Each of these indirect paths was significant (see Table 16).

9 Discussion

Consistent with study 1, this study provides evidence that response entropy (RE) taps a distinct latent construct from other measures of careless responding. In this study, a factor analysis supports a latent construct comprised of only the RE and LS facets. Correlations and factor analysis also distinguish RE from response styles, whereas other measures of CR were not distinguished as such; for example, the regression type measures factored with social desirability and acquiescence and within-person standard deviation (WS) factored with midpoint and extreme responding. Person-fit type indices also, again with Mahalanobis distance (MD), comprised their own factor; correlations however do show that MD, the number of Guttman errors (G_n^p), and standardized log-likelihood (L_z^p) are all impacted by extreme score response styles.

Additionally, consistent with study 1, RE was not related to participant race. LS was the only other measure of CR unrelated to participant race. Previous research has demonstrated that response styles to surveys vary by culture (Bachman & O'Malley, 1984; He et al., 2014; He & Van De Vijver, 2015; Hui & Triandis, 1989; Smith, 2004). It was hypothesized that differences between Asian and White participants' scores on CR indices would best be accounted for by acquiescent style responding, and that differences between Black and White participants' scores on CR indices would best be accounted for by extreme style responding. These hypotheses were partially supported.

Differences between Asian and White participants on MD, PS, and EO were mediated by acquiescent responding, and these findings converge with research demonstrating Asian participants to have a more acquiescent response style than Whites (Smith, 2004). Acquiescent responding was also related to lower G_n^p scores, although the direct effect between participant race and G_n^p was not significant. These findings suggest that MD, PS, EO, and G_n^p in particular may tend to conflate Asians' culturally distinct responding style with careless responding. Although direct (marginal) differences between Asian and White participants' scores on PA were also observed, a cultural response style explaining these differences was not discovered. Future research may wish to explore other explanatory mechanisms for differences in PA scores among Asian and White participants, particularly given that the PA emerged among other metrics as a particularly strong indicator of careless responding. For example, one possibility is that differences in patterns of endorsement emerged on the beliefs about race scale because of Asian Americans more complex and variant experiences with race and racism compared to White/ European Americans (e.g. Lin, Kwan, Cheung, & Fiske, 2002). As one word of caution, in the results section, the factor analysis placed acquiescent responding in the same factor with PS and EO, thus, the indirect findings related to this variable are somewhat redundant. Nonetheless, the factor analysis also points to the conflation of these metrics of careless responding (PS and EO) and culturally distinct responding.

Differences between Black and White participants on MD, EO, and WS were mediated by midpoint responding, and these findings partially converge with previous research

Table 15
Direct and indirect effects of race (asian vs. White) on careless response indices

	Race-effect	S.E.	95% C.I. Lower	95% C.I. Upper
Mahalanobis Distance				
Direct	−0.23*	0.12	−0.46	−0.01
Thru midpoint	0.02	0.02	−0.02	0.08
Thru extreme	0.09	0.07	−0.04	0.23
Thru asquiescent	−0.07	0.04	−0.17	−0.01
Thru social des.	0.01	0.02	−0.03	0.06
Psychometric Synonym Index				
Direct	0.26	0.14	−0.01	0.53
Thru midpoint	0.05	0.05	−0.02	0.17
Thru extreme	0.02	0.03	−0.02	0.08
Thru asquiescent	0.10*	0.04	0.02	0.19
Thru social des.	0.00	0.03	−0.05	0.06
Psychometric Antonym Index				
Direct	−0.33	0.19	−0.70	0.04
Thru midpoint	0.00	0.04	−0.09	0.07
Thru extreme	−0.04	0.05	−0.15	0.04
Thru asquiescent	−0.05	0.05	−0.17	0.04
Thru social des.	0.08	0.06	−0.20	0.01
Even-Odd Index				
Direct	0.27	0.14	−0.01	0.55
Thru midpoint	0.05	0.05	−0.04	0.15
Thru extreme	−0.02	0.02	−0.07	0.02
Thru asquiescent	0.14*	0.06	0.04	0.26
Thru social des.	0.04	0.04	−0.03	0.12
Response Entropy Index				
Direct	−0.08	0.16	−0.39	0.23
Thru midpoint	0.05	0.05	−0.04	0.17
Thru extreme	−0.07	0.06	−0.21	0.02
Thru asquiescent	0.03	0.04	−0.05	0.12
Thru social des.	0.05	0.04	−0.03	0.14
Within-person Standard Deviation				
Direct	−0.02	0.10	−0.22	0.17
Thru midpoint	0.05	0.04	−0.04	0.12
Thru extreme	0.11	0.08	−0.04	0.27
Thru asquiescent	−0.03	0.03	−0.10	0.02
Thru social des.	0.01	0.02	−0.03	0.05
Normed Guttman Errors				
Direct	0.11	0.11	−0.12	0.33
Thru midpoint	0.01	0.02	−0.01	0.06
Thru extreme	0.13	0.10	−0.07	0.32
Thru asquiescent	−0.04*	0.03	−0.12	−0.00
Thru social des.	0.05*	0.03	−0.11	−0.00
Standardized Log-Likelihood				
Direct	0.17	0.14	−0.10	0.44
Thru midpoint	0.00	0.02	−0.04	0.03
Thru extreme	−0.08	0.06	−0.22	0.03
Thru asquiescent	−0.03	0.03	−0.08	0.02
Thru social des.	0.03	0.03	−0.02	0.10

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table 16
Direct and indirect effects of race (Black vs. White) on careless response indices

	Race-effect	S.E.	95% C.I. Lower	95% C.I. Upper
Mahalanobis Distance				
Direct	−0.55**	0.12	−0.78	−0.31
Thru midpoint	−0.04*	0.03	−0.10	0.00
Thru extreme	−0.04	0.06	−0.18	0.08
Thru asquiescent	−0.01	0.03	−0.09	0.06
Thru social des.	0.02	0.02	−0.01	0.07
Psychometric Synonym Index				
Direct	0.34**	0.13	0.09	0.59
Thru midpoint	−0.07	0.05	−0.18	0.02
Thru extreme	−0.01	0.02	−0.06	0.02
Thru asquiescent	0.01	0.03	−0.05	0.09
Thru social des.	−0.01	0.02	−0.05	0.02
Psychometric Antonym Index				
Direct	0.06	0.18	−0.29	0.41
Thru midpoint	0.01	0.03	−0.04	0.08
Thru extreme	0.00	0.02	−0.04	0.05
Thru asquiescent	0.00	0.02	−0.05	0.05
Thru social des.	−0.04	0.04	−0.12	0.02
Even-Odd Index				
Direct	0.25*	0.13	0.00	0.50
Thru midpoint	−0.08*	0.05	−0.18	0.00
Thru extreme	0.00	0.02	−0.03	0.03
Thru asquiescent	0.02	0.04	−0.08	0.10
Thru social des.	−0.01	0.02	−0.05	0.03
Response Entropy Index				
Direct	−0.11	0.15	−0.40	0.17
Thru midpoint	−0.05	0.05	−0.17	0.02
Thru extreme	0.02	0.04	−0.05	0.12
Thru asquiescent	0.00	0.02	−0.04	0.04
Thru social des.	0.02	0.02	−0.03	0.07
Within-person Standard Deviation				
Direct	−0.25**	0.09	−0.43	−0.06
Thru midpoint	0.08*	0.04	−0.15	−0.00
Thru extreme	0.05	0.09	−0.24	0.11
Thru asquiescent	0.00	0.02	−0.05	0.03
Thru social des.	0.01	0.01	−0.01	0.04
Normed Guttman Errors				
Direct	−0.33**	0.10	−0.53	−0.13
Thru midpoint	−0.02	0.02	−0.07	0.02
Thru extreme	−0.08	0.09	−0.26	0.05
Thru asquiescent	−0.01	0.03	0.07	−0.00
Thru social des.	−0.01	0.01	−0.04	0.02
Standardized Log-Likelihood				
Direct	0.55**	0.13	0.30	0.80
Thru midpoint	0.01	0.03	−0.04	0.07
Thru extreme	0.03	0.05	−0.07	0.14
Thru asquiescent	0.00	0.031	−0.02	0.02
Thru social des.	−0.01	0.02	−0.05	0.03

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

demonstrating Black participants to have more extreme score responding, and presumably lower midpoint responding than Whites (Bachman & O'Malley, 1984). It is curious that midpoint rather than extreme responding emerged as the operative mechanism in the parallel mediation analysis, however, given their shared variance as inverse constructs it is unlikely both would emerge as unique predictors. Although direct differences between Black and White participants' scores on PS, G_n^p, and L_z^p , were also observed, a cultural response style explaining these differences was not discovered. Again, the factor analysis placed both midpoint responding and WS in the same factor, thus, this indirect finding related to this variable is somewhat redundant. Regardless, if one were to use WS to measure CR, one precaution is that this measure is more likely to flag members of racial and ethnic groups (i.e., Blacks and Latino/as; who tend to use more variability (i.e., less midpoint and more extreme scores) in their responding styles. This raises an interesting question about why then the RE index-also a variability type measure - is not likely to conflate culturally diverse and careless responding. The answer is that while RE does measure variability, it determines proportions of responses without weighing the actual values of those responses. For instance, a person who responds to four items with 1,1,6,6 would resemble an extreme responding pattern and would have a much higher WS score than someone who responded 3,3,4,4. However, their RE score would be the same since both participants endorsed two different options two times each. While these initial findings suggest that the RE is a good option for researchers seeking to determine CR without inadvertently removing cultural variability from data sets due to confounding cultural values with CR, considerably more research applying the RE to various sociocultural group (e.g., based on nationality, ethnicity, gender, and social class) is needed.

10 General Discussion and Conclusion

The response entropy index appears to be a viable option for detecting potentially careless response patterns, and two studies suggest that it appears to be tapping a relatively distinct latent construct from other previously established regression type and person-fit type measures. Following study 1, I recommended that for best practices in screening for careless responses, researchers should select one metric from each of the three classes: variability, regression, and person-fit type. In fact, to aid researchers with application of these metrics, it may be possible to develop a single algorithm that can automatically draw from metrics from each of these classes to award "points" to an overall carelessness score. Leiner (2019) proposed such an algorithmic method to detecting low quality data by awarding points to patterning that could reflect carelessness; as examples, points were awarded for instances of repeated value endorsements (i.e., overly repetitive) and also when respondents began creating diagonal lines with their responses (i.e., overly scattered). Although Leiner (2019) found only minimal support for their algorithm for detection of low quality data, the idea is a promising one and should be considered by future programmers and researchers, again, perhaps with the goal of combining variability, regression, and person-fit type indices.

Specifically I recommended: the RE index which emerged as the strongest predictor in its class for detecting BS with human careless data and computer-uniform data; 2) the PA index which emerged in the multinomial regression as a significant predictor of all three types of BS data; and 3) the L_z^p which emerged as the strongest predictor in its class for detecting BS with human careless data and computer-uniform data. After examining the potential for conflation of culturally diverse responding and careless responding in study 2, I would continue to recommend these three metrics. Although some conflation effects were discovered for L_z^p and PA these effects were relatively innocuous compared to other metrics. Across two studies, no demographic differences were determined for the RE index and thus this metrics should be relatively safe regarding conflation effects.

It is curious that human BS responders tended to opt to use overly consistent responding rather than an overly scattered approach, although intuitively this choice makes sense as it may be faster than a more complex patterning[11]. Thus the findings in this study regarding the effectiveness of the RE index for detection of careless human data could be an artifact of the instructions in which BS responders were permitted to falsify data. Without such permission human BS responders, perhaps hoping to go undetected, may opt for a more complex and scattered patterning. A future research project might consider revising the task instructions or even develop experimental manipulations to encourage more overly scattered human BS responses. Relatedly, the linear statistics employed in the current analyses may have underappreciated the predictive capacity of the RE index. For example, in the logistic regression analysis, the RE index could only work in one direction for each dataset given the binary nature of the outcome variable (BS vs. non-BS); in the human careless dataset it worked to predict overly consistent responding, in the computer uniform dataset it worked to detect overly scattered responding. This means, for example, that in the analysis of human careless data, overly scattered responses loaded into the non-BS bin, and thus did not contribute to the predictivity of BS. A future research project may wish to devise a multi-level outcome variable in order to employ non-linear analyses of the predictive capacity of the RE index. To be clear, although the use of binary out-

[11] A post hoc frequency analysis of z-scores for REI values in study 2 determined that 41 and 13 participants fell within 10th and 90th percentiles respectively; thus, while the majority of questionable protocols were overly consistent, there were some instances of overly scattered responding.

come variables limited the capacity of the RE index to detect either overly consistent or overly repetitive responses within each dataset used in this study, in practical application the RE index is capable of flagging both types within a single dataset. As mentioned in the author note, a user-friendly response entropy index calculator is available online at https://sagenm.shinyapps.io/REICalculator/ and is free for researchers to use to screen their data for careless responders.

Although I found some evidence for underlying latent constructs, factor analyses in both studies suffered from moderate fit statistics; thus, interpretation of the structural regression analyses was limited and I primarily relied upon the multinomial logistic regression for interpretation. Furthermore, some of the findings related to the factor analyses were not always clear, for example, why MD clustered with the person-fit type and why social desirability and acquiescent responding clustered with regression type indices. Future research should continue a line of research (e.g. Grau et al., 2019)huang12meade12 attempting to understand the latent constructs measured by proposed indices of CR.

In addition to proposing a new index of careless responding, these analyses were unique in that they determined CR from measures that are more typical of social or sociocognitive rather than clinical psychological research (Worthington & Whittaker, 2006). The previous studies reviewed above have most commonly derived CR from measures of personality (Huang et al., 2012; Johnson, 2005; Marjanovic et al., 2015; Meade & Craig, 2012). In fact, one could raise issue with the fact that both scales chosen from which to derive measures of CR (CoBRAS and BARS) were both related to perspectives on race and thus content endorsement is likely to be particularly effected by participant race. This is certainly true although no different than any other measures (e.g., social, vocational, cognitive, behavioral, etc.) in which participant race differences occur. Additionally, as a reminder, the indices of CR examined in this study are "content independent" (Huang et al., 2012) by nature. How degree of endorsement of content or proficiency on a measure relates to careless responding is an interesting question in its own right, one that has begun to be taken up by researchers. Rios, Guo, Mao, and Liu (2017) for example, have examined how ability on intelligence tests relates to CR. This question is however, beyond the scope of the current study.

Not only are the scales used in this study different in content than previous research, they also tend to be considerably shorter than the personality inventories and tests of intellectual ability used in previous studies. For instance the personality databases used by many CR researchers (Huang et al., 2012; Johnson, 2005; Meade & Craig, 2012) comprised a total of 300 total items, although exact calculation of CR metrics incorporated different numbers of items. For example, the psychometric antonym index has been computed from the 30 most strongly negatively related item pairs (Huang et al., 2012) and also from just the five item pairs negatively correlating at least as strongly as −0.60 (Meade & Craig, 2012). Other studies have determined CR indices from a 72 item measure of intellectual ability (Rios et al., 2017), a 60 item short form personality inventory (Marjanovic et al., 2015), a 31 item quality of life scale (Schneider et al., 2017), and simulated data sets with scales of 30 items (Rios et al., 2017). Thus, the 20 and 16 item scales used in this study are comparably smaller and findings may be only specific to measures of this size. The uniqueness of the content (i.e., social psychological) and item length contributes to our understanding of the generalizability of the effectiveness of previously used measures of CR to very short form scales. One challenge with very short form scales arose in the computation of regression type indices with limited numbers of item pairs which could not be computed when there was no variability in scores on one side of the equation. Niessen et al. (2016) also raised the issue of the use of regression type indices with data comprised from smaller item numbers and recommended in these instances the use of person-fit type indices. In the case of the RE index, much more research is now needed to examine if its effectiveness can generalize beyond short form scales.

Perhaps the biggest question looming among CR research is how to practically implement these indices. By now, a considerably sized body of research offers multiple, relatively effective methods for detecting CR but little consensus on how to implement them. Some authors have weighed in on the idea of implementation of a "zero-tolerance" cutoff; in this practice one simply removes data that exceeds a set percentile on a specified CR index. Huang et al. (2012) demonstrated that with a sample of 345 college students, even removal of the top 1% of careless responders in the sample could significantly improve the psychometric properties of the scale (e.g., Cronbach alphas). Others are more cautious about the idea of zero-tolerance cutoffs because of the potential for removing true variability in the dataset, and certainly findings from study 2 raise caution related to the idea of blindly removing data given that less modal populations such as racial and ethnic minorities may be more likely than White participants to have their attentive data improperly removed. Some possible solutions include an error-balancing protocol in which the likelihood of false positives is balanced with the likelihood of false negatives (see D. S. Kim et al., 2018), and the recommendation by Niessen et al. (2016) to pre-register uses of CR indices including pre-specified cutoffs.

For the RE index specifically, should future researchers adopt this measure, overtime relatively objective "cutoffs" for extreme high and low scores may emerge. In the "true" human dataset in study 2, the range of RE scores was 0.00 to 0.76 with an average of 0.53 and a standard deviation of 0.15. Based on a normal distribution principle, 10^{th} and 90^{th}

percentile extreme scores would occur at 1.28 standard deviations above and below the mean; thus preliminary guidelines from study 2 would be that scores below 0.34 and above 0.72 should raise suspicion. Application of the RE index to many more samples would be needed to begin to confirm such guidelines. Moreover, standardizing of high and low RE index values would need to be tested across scales of varying ranges of item response options; in all of the datasets used in this paper, scales ranged from 1 to 6. Theoretically, the RE index can be applied to scales with broader response option ranges (e.g., 1 to 10) without artificially inflating values due to the confounding of response type and response value, yet this claim remains to be empirically established.

On the other hand, there may be an argument for not developing objective cutoffs, but rather to tailor RE index implementation to each specific dataset. For example, in this study the RE index was more effective at distinguishing uniform random data compared to normally distributed random data from true response sets; the platykurtic distribution of the uniform data resulted in more scatter across item responses whereas the normal distribution more closely resembled true response sets which tend to cluster around a participant's particular item response tendency (e.g., 2's and 3's). In other words, the effectiveness of RE index may depend on the extent to which careless data distribute notably differently than the true data. Yet, what happens in instances where *true data* distributions are skewed? For example, one could imagine a self-reported chronic pain scale; when administered to a sample of healthy young adults, modal item responses may be primarily comprised of the lowest possible item score (i.e., a series of 1's) resulting in RE scores frequently occurring below the objective low threshold (e.g., 0.34). In this instance, RE index may be more effective if applied relative to the specific dataset at hand, for example, by converting RE index scores to z-scores and flagging z-scores of 2 or greater.

Given that these questions can only be addressed by much further research, my final recommendation would be to continue to actively consider all of these options but not rush to standardize a protocol for data screening, this research is still relatively young and much more is to be learned. I would encourage other researchers to continue to evaluate the response entropy index as a potentially viable option.

Acknowledgement

All datafiles and syntax for constructing indices and running analyses can be accessed through the Open Science Framework [follow link: OSF Response Entropy Index]. In addition, a user-friendly response entropy index calculator is available online at https://sagenm.shinyapps.io/REICalculator/ and is free for researchers to use to screen their data for careless responders. Special thanks to Sage Mahannah for development of the calculator and Dr. Amanda Montoya (University of California, Los Angeles) and Dr. Holly Laws (University of Massachusetts, Amherst) for statistics consultation.

References

Bachman, J. G., & O'Malley, P. M. (1984). Black-white differences in self-esteem: Are they affected by response styles? *American Journal of Sociology*, *90*(3), 624–639.

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on the MMPI–A. *Journal of personality assessment*, *68*(1), 139–151.

Beach, D. A. (1989). Identifying the random responder. *The Journal of psychology*, *123*(1), 101–103.

Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological assessment*, *4*(3), 340.

Böckenholt, U. (2013). Modeling multiple response processes in judgment and choice. *Decision*, *1*(S), 83–103. doi:10.1037/2325-9965.1.s.83

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, *111*(2), 218.

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., . . . Nutt, D. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience*, *8*, 20.

Costa Jr, P. T., & McCrae, R. R. (2008). *The revised neo personality inventory (neo-pi-r).* Sage Publications, Inc.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, *10*(1), 7.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. doi:10.1037/h0047358

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological methods*, *1*(1), 16.

Daprati, E., Sirigu, A., Desmurget, M., Martinelli, E., & Nico, D. (2019). Willingness towards cognitive engagement: A preliminary study based on a behavioural entropy approach. *Experimental brain research*, *237*(4), 995–1007.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of personality and social psychology*, *48*(1), 82.

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

Hayes, A. F. [A. F.]. (2013). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. New York, NY: Guilford Press.

He, J., de Vijver, F. J. V., Espinosa, A. D., & Mui, P. H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles. *International Journal of Cross Cultural Managemen*, *14*(3), 306–322. doi:10.1177/1470595814541424

He, J., & Van De Vijver, F. J. (2015). Effects of a general response style on cross-cultural comparisons: Evidence from the teaching and learning international survey. *Public Opinion Quarterly*, *79*(S1), 267–290.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

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.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of cross-cultural psychology*, *20*(3), 296–309.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological assessment*, *15*(4), 446.

Jackson, D. L., Gillaspy Jr, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological methods*, *14*(1), 6.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, *39*(1), 103–129.

Kim, B. S., Atkinson, D. R., & Yang, P. H. (1999). The asian values scale: Development, factor analysis, validation, and reliability. *Journal of counseling Psychology*, *46*(3), 342.

Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A., & King, K. M. (2018). Detecting random responders with infrequency scales using an error-balancing threshold. *Behavior research methods*, *50*(5), 1960–1970.

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment*, *76*(2), 315–332.

Lang, J. W., Lievens, F., De Fruyt, F., Zettler, I., & Tackett, J. L. (2019). Assessing meaningful within-person variability in likert-scale rated personality descriptions: An IRT tree approach. *Psychological Assessment*, *31*(4), 474.

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229–248.

Lin, M. H., Kwan, V. S. Y., Cheung, A., & Fiske, S. T. (2002). Stereotype content model explains prejudice for an envied outgroup: Scale of anti-asian american stereotypes. *Personality and Social Psychology Bulletin*, *31*(1), 34–47.

Mahalanobis, P. C. (1936). *On the generalized distance in statistics*. National Institute of Science of India.

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.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. doi: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. doi:10.1177/1073191115577800

Neville, H. A., Lilly, R. L., Duran, G., Lee, R. M., & Browne, L. (2000). Construction and initial validation of the color-blind racial attitudes scale (CoBRAS). *Journal of counseling psychology*, *47*(1), 59.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*(2), 330–357. doi:10.1037/0033-295x.109.2.330

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based question-

naires: Which method to use? *Journal of Research in Personality*, *63*, 1–11.

Preacher, K. J., & Hayes, A. F. [Andrew F.]. (2008). Contemporary approaches to assessing mediation in communication research. In XXXX (Ed.), *The sage sourcebook of advanced data analysis methods for communications research* (pp. 13–54). doi:10.4135/9781452272054.n2

Quillian, L., & Redd, R. (2009). The friendship networks of multiracial adolescents. *Social Science Research*, *38*(2), 279–295.

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.

Roma, P., Mazza, C., Mammarella, S., Mantovani, B., Mandarelli, G., & Ferracuti, S. (2019). Faking-good behavior in self-favorable scales of the MMPI-2. *European Journal of Psychological Assessment*.

Schneider, S., May, M., & Stone, A. A. (2017). Careless responding in internet-based quality of life assessments. *27*(4), 1077–1088. doi:10.1007/s11136-017-1767-2

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. sage.

Simel, D. L., Easter, J., & Tomlinson, G. (2013). Likelihood ratios, sensitivity, and specificity values can be back-calculated when the odds ratios are known. *Journal of clinical epidemiology*, *66*(4), 458–460.

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology*, *35*(1), 50–61. doi:10.1177/0022022103260380

Sue, D. W., & Sue, D. (2012). *Counseling the culturally diverse: Theory and practice, 6th ed.* Hoboken, NJ: John Wiley and Sons Inc.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics*. Pearson Boston, MA.

Tawa, J. (2017a). Asymmetric peer selections among blacks, asians, and whites in a virtual environment: Preliminary evidence for triangulated threat theory. *The Journal of social psychology*, *157*(6), 736–753.

Tawa, J. (2017b). The beliefs about race scale (BARS): Dimensions of racial essentialism and their psychometric properties. *Cultural Diversity and Ethnic Minority Psychology*, *23*(4), 516.

Tukey, J. W. (1977). Exploratory data analysis.

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.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. doi:10.1177/0011000006288127