

# Detecting random responders with infrequency scales using an error-balancing threshold

Dale S. Kim<sup>1</sup> · Connor J. McCabe<sup>2</sup> · Brianna L. Yamasaki<sup>2</sup> · Kristine A. Louie<sup>2</sup> · Kevin M. King<sup>2</sup>

Published online: 21 September 2017  
© Psychonomic Society, Inc. 2017

**Abstract** Infrequency scales are becoming a popular mode of data screening, due to their availability and ease of implementation. Recent research has indicated that the interpretation and functioning of infrequency items may not be as straightforward as had previously been thought (Curran & Hauser, 2015), yet there are no empirically based guidelines for implementing cutoffs using these items. In the present study, we compared two methods of detecting random responding with infrequency items: a zero-tolerance threshold versus a threshold that balances classification error rates. The results showed that a traditional zero-tolerance approach, on average, screens data that are less indicative of careless responding than those screened by the error-balancing approach. Thus, the de facto standard of applying a “zero-tolerance” approach when screening participants with infrequency scales may be too stringent, so that meaningful responses may also be removed from analyses. Recommendations and future directions are discussed.

**Keywords** Random responding · Careless responding · Infrequency scales · Data cleaning

Survey studies often presume that participants are attentive and responding conscientiously. However, there is increasing

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.3758/s13428-017-0964-9>) contains supplementary material, which is available to authorized users.

---

✉ Dale S. Kim  
dalekim25@ucla.edu

<sup>1</sup> Department of Psychology, University of California, Box 951563, 1285 Franz Hall, Los Angeles, CA 90095-1563, USA

<sup>2</sup> Department of Psychology, University of Washington, Seattle, Washington, USA

evidence that this is not necessarily the case (Curran, 2016; Maniaci & Rogge, 2014; Meade & Craig, 2012). Estimates of the rate of careless or inattentive responding vary from as low as 3%–9% (Maniaci & Rogge, 2014) to 46% (Oppenheimer, Meyvis, & Davidenko, 2009) and may depend on the demographic (Berinsky, Margolis, & Sances, 2014). When participants are providing careless responses, data quality suffers by the introduction of systematic or nonsystematic variance (Maniaci & Rogge, 2014; Meade & Craig, 2012) and may obscure the ability to detect effects or spuriously introduce them (Credé, 2010; Huang, Liu, & Bowling, 2015). For example, the inclusion of careless responders has been found to cause failures to replicate established experimental findings (Oppenheimer et al., 2009; Osborne & Blanchard, 2011), diminished correlations between constructs (Fervaha & Remington, 2013) and also influence the model selection of psychological factor models (Woods, 2006). Given the increasing reliance on computerized and online surveys, which may be at increased risk for careless responding, improving data screening methods is critical in assuring analysis quality.

## Types of careless responding

A participant may respond carelessly on a survey in several ways. Examples include uniformly random responding (selecting each possible response with roughly equal probability), long-string responding (selecting the same response over a large number of items; Johnson, 2005) or even nonrandom patterned responding (e.g., repeating an ascending sequence of numbers). The common element between these response patterns is the lack of responding to the item content. This has also been described as content nonresponsivity (e.g., Nichols, Greene, & Schmolck, 1986) or inattentive responding (e.g., Johnson, 2005). It should be noted that it is not necessary that

a participant be strictly content responsive or nonresponsive as a whole, but may display some degree of severity between the two extremes. For example, participants who are hurriedly responding may only be responsive to some, but not all, of the item's content (e.g., Berinsky et al., 2014).

For the present study, we focus on uniformly random responding. There are several reasons for this. First, careless responding is of particular interest due to its propensity to manifest in low/medium stakes testing, in which participants may not necessarily be fully motivated to provide sufficient effort or attention (Curran, 2016). This is a ubiquitous research situation in psychological studies, especially with student samples, as well as in large scale epidemiological samples (such as the annual Monitoring the Future survey or the Youth Risk Behavior Surveillance System; Brener et al., 2013; Johnston, O'Malley, Miech, Bachman, & Schulenberg, 2016). Additionally, prior research has found that uniformly inconsistent responding is by far the most common type of careless responding and benefits from the ease of detection (Meade & Craig, 2012).

It is worth clarifying the exact meaning of the term “uniformly random responding” for the purposes of the present study. In a mathematical sense, the term “uniform distribution” refers to a situation in which all possible outcomes of a random variable are equally probable. Rolling a fair die or flipping an unbiased coin are common examples. In the case of random responding however, research has shown that human behavior tends not to mimic random variables exactly, even when it is attempted to do so (Bakan, 1960; Tune, 1964). Thus, we will make a distinction between uniformly random responding and a uniformly random variable. We will refer to uniformly random responding (henceforth simply as random responding) as a type of content nonresponsive responding in which each response is selected with *roughly* equal probability or simply selecting from a variety of response options. In contrast, a uniform random variable will refer to a more stringent stochastic process, in which the probability of each response is *strictly* equal, and each draw is independent and identically distributed.

## Infrequency scales

Several methods exist to screen for random responding, which include post hoc methods (e.g., consistency, outlier, response time analysis) and/or specially constructed scales inserted into the test or survey (e.g., self-report item engagement; Berry et al., 1992). One common type of scale is an infrequency scale, which consists of items with highly or absolutely skewed response distributions. This skew stems from the fact that these items are written such that there is a distinct set of plausible or correct responses as well as a distinct set of unlikely or incorrect responses to complement. For example, the item “I am

answering a survey right now,” with five response options ranging from *Strongly Disagree* to *Strongly Agree* would have at least two clear answers (*Agree* or *Strongly Agree*). Thus, if one assumes that a respondent has in fact read the item stem and responded conscientiously, any other response would be *infrequently* (or never) chosen. By this logic, those who choose infrequent responses are presumed to be responding carelessly (at least at that particular point of time).

The use of infrequency scales have a long history, dating back to at least the Minnesota Multiphasic Personality Inventory (MMPI), which utilized infrequency items measuring tendencies toward responding in an overly favorable manner (“L” scale, 15 items) as well as exaggerating symptoms of psychopathology (“F” scale, 60 items; Hathaway & McKinley, 1951). Although these types of scales have been labeled differently over time, such as bogus items (Meade & Craig, 2012), conscientious-responding scales (Marjanovic, Struthers, Cribbie, & Greenglass, 2014), or random-responding scales (Beach, 1989), they are all based on the infrequency technique of data validation.

It should be noted that infrequency scales may have a different interaction with other invalid, yet noncareless, response types. Specifically, infrequency scales may not capture noncareless invalid responses (e.g., lying or faking) unless special care is made to design the item for that purpose (e.g., MMPI). However, one sub-class of noncareless responses that infrequency scales may capture are mischievous responses. Mischievous responding occurs when respondents are intentionally providing extreme and often untruthful responses out of self-amusement (Robinson-Cimpian, 2014). In other words, certain infrequency items that employ content that can be construed as extreme, odd, or entertaining (e.g., “My main interests are coin collecting and interpretive dancing”; Maniaci & Rogge, 2014) may also be screening mischievous responders in addition to careless responders.

## Infrequency scale implementation

In recent research, investigators who have reported using infrequency scales tended to take a zero-tolerance approach when screening data, which excludes any respondents who have one or more incorrect responses to infrequency items (e.g., Fervaha & Remington, 2013; Osborne & Blanchard, 2011; Periard & Burns, 2014). Although this approach may seem reasonable, it makes several assumptions about those respondents. First, the zero-tolerance approach assumes that any invalid response to an infrequency item means that the respondent was always invalidly responding. Yet it is possible that careless responding arises due to state factors in which participants momentarily respond carelessly (e.g., environmental

distraction), as opposed to a trait-like tendency toward carelessness.

Second, zero tolerance assumes that the false positive detection rate of random responding is zero. However, this may depend on the nature of the infrequency items being utilized. For example, consider an infrequency item that reads “I drink more than ten glasses of milk a day” with seven response options from *Strongly Disagree* to *Strongly Agree*. Typically, conscientious responders would not endorse this item, as drinking ten glasses of milk a day is extremely unlikely. However, for a small proportion of conscientious responders this item may actually be true. Alternatively, a proportion of conscientious responders may reply *Neutral* or *Slightly Disagree* because they drink large amounts of milk, though not ten glasses. Although in practice this may yield a negligibly small false-negative rate for a single item, these small probabilities will accumulate when taken across multiple items, increasing the chances of incorrectly flagging a conscientious responder.

Indeed, previous research supports this notion. Curran and Hauser (2015) interviewed respondents after administering infrequency scales to gain insight into how these items can be interpreted. In one example, conscientious responders gave slight endorsement or neutral responses to the item “I am paid biweekly by leprechauns” because they are paid biweekly, but not by leprechauns. In another example, Meade and Craig (2012) found that the item “I have never spoken to anyone who was listening” was endorsed by a large amount of their sample, and thus it may not have been interpreted as intended. Therefore, methods that are more integrative will attempt to differentiate between idiosyncratic interpretations, and stable patterns of careless responding across multiple tests or instances. Altogether, if these assumptions are not carefully considered in screening, a subset of highly conscientious responders may remain in the data as a result, which could exaggerate effects or otherwise cause the results to be distorted.

**Error-balancing threshold** Thus, to accommodate the fact that infrequency items do not always function as intended, we propose the usage of a technique that balances true positive and false positive error rates. This idea is based in receiver operating characteristic (ROC) analysis, in which true positive rates (also called sensitivity) are balanced against false positive rates (calculated as  $1 - \text{specificity}$ ) to optimize classification performance (Fawcett, 2006). Plotting these points in a two-dimensional space yields a useful diagnostic plot, called a ROC curve. We propose using the cutoff that balances these two quantities equally, which is also known as the Youden index (Youden, 1950) in ROC literature.

In practice, the actual true positive rates and false positive rates of random responding are not known. Therefore, to adapt ROC classification to infrequency scales, we use practical

approximations for these quantities. For true-positive rates, we propose that the cutoffs are mapped such that they reflect the probability of a uniform random variable to approximately model random responding. A uniform random variable should approximate the inconsistent nature of careless responding and serve as a practical metric of evaluation consistent with prior research (e.g., Marjanovic et al., 2014). The mathematical considerations needed to calculate these cutoffs are described with probability functions below. To serve as a proxy for false positive rates, we propose the usage of the proportion of data screened, due to the monotonic relationship between the two. That is, as the amount of screened data increases, the possibility of a false positive can only increase as well. Furthermore, the proportion of the data removed at each cutoff point serves as a practical diagnostic regarding the nature of the sample as well.

**Probability functions** We utilize probability functions to calculate the true positive detection rate of the infrequency scale. Using a uniform random variable to model random responses, the probability of choosing an incorrect response on any infrequency item will be distributed as a Bernoulli random variable with parameter  $p_i$ , where  $p_i$  is the proportion of responses considered incorrect with respect to the total number of responses within an item. If we further suppose that all infrequency items are identically distributed (where all  $p_i = p$ ), then the sum of incorrect responses is distributed as a binomial random variable. If we let  $X$  represent the number of questions answered incorrectly, then the probability of each outcome of  $X$  is well known to be given by

$$p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

This formulation is convenient for cases in which all questions have identical response sets and number of correct/incorrect responses. However, it may not necessarily be the case that all questions will have identical response sets. For example, a researcher may want to use multiple scales and camouflage infrequency items by matching response sets between the infrequency item and the items of the surrounding scale. When using non-identical response sets, having infrequency items with differing Bernoulli distributions will be likely. In such a case, the binomial distribution will be an improper and inaccurate probability model. Instead, the sum of nonidentically distributed Bernoulli random variables is modeled by the Poisson binomial distribution. A succinct formulation is provided by Wang (1993) as follows:

$$p(X = k) = \sum_{A \in F_k} \prod_{i \in A} p_i \prod_{j \in A^c} (1-p_j),$$

where  $A^c$  is the complement of  $A$ , and  $F_k$  is the subset of all integer combinations that can be selected from the set  $\{1, 2, 3,$

$\dots, n\}$ . Thus, if  $n = 4$ , then  $F_3 = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}\}$ . Although the derivation of this expression is beyond the scope of this study, interested readers are encouraged to see Wang (1993) for a rigorous treatment of the subject. We wrote a software implementation of this function in the R language (R Development Core Team, 2015) and included it in the [Online Appendix](#).

The main advantage of the Poisson binomial distribution is that it takes into account all permutations of differing Bernoulli parameters. This gives the researcher the flexibility to use different response sets between their infrequency items or to create response sets with different response criteria. For example, one infrequency item can have a .75 probability of flagging a uniform random variable, in which another question can have a .80 probability. In these cases, the Poisson binomial distribution models the number of flagged responses accurately. Note once again that when all Bernoulli parameters are identical, the Poisson binomial distribution reduces to the classic binomial distribution.

**Point of error balance**<sup>1</sup> Once the proper probability distribution is constructed, the threshold of error balance can be calculated. The objective is to choose a cutoff point such that the probability of a uniform random variable surviving the flagging process and the proportion of data being eliminated are balanced equally. This can be represented mathematically with the following expression:

$$\underset{x}{\operatorname{argmax}} \{P(x) + F(x) - 1\},$$

where  $x$  is the number of incorrectly answered infrequency items,  $F(x)$  is the proportion of data that would be retained if the cutoff point  $x$  or greater is used and  $P(x)$  is the probability of a uniform random variable realizing the outcome of  $x$  or greater. Once again, this is analogous to ROC analysis in which  $P(x)$  represents the sensitivity,  $F(x)$  represents the specificity, and the point of error balance represents the aforementioned Youden index. Equivalently, the point of error balance is the cutoff point with the minimum Euclidean distance between itself and the point (0, 1) on a ROC curve (Schisterman, Perkins, Liu, & Bondell, 2005).

## Study aims

Although infrequency scales as a whole benefit from their ease of administration (Meade & Craig, 2012), there is little empirical research on exactly what thresholds should be implemented or what effects different thresholds can have on data quality. If data screening is not employed at

all, careless responders might be adding extraneous variance that reduces statistical power. On the other extreme, using an overly stringent threshold (such as zero tolerance, which we found the majority of studies utilize) may ultimately result in the removal of meaningful responses as well. Thus, the goal of the present investigation was to empirically test the widely used zero-tolerance threshold against the proposed error-balancing threshold, and to examine the indication of random responding associated with each. This is achieved through two studies. First, we conducted a simulation study to validate and demonstrate the error-balancing threshold under several hypothetical scenarios (Study 1). Second, we applied the error-balancing and zero-tolerance thresholds to two independent datasets and evaluated their performance (Study 2).

## Study 1

### Method

We designed a simulation to emulate a realistic data collection scenario. Simulated responses for ten infrequency items and ten substantive variables were generated. There was a total of  $n = 1,000$  responses, allocating 900 (90%) of these to be simulated conscientious responses and the remaining 100 (10%) to be simulated random responses. For conscientious responses, the substantive variables were simulated using a multivariate normal distribution, with means and variance set to 1 and with covariances set to .3. Once these variables were generated, they were rounded to the nearest integer with a floor and ceiling set to 1 and 5, respectively. The purpose of the rounding step was to create skewed ordinal data typically found in behavioral survey measures. For the random responses to the substantive variables, a discrete uniform distribution ranging from 1 to 5 was used.

To generate responses to infrequency items, we used a simple Bernoulli distribution. The probability of a random response being flagged by an infrequency item was set to .8. This emulates a scenario in which the infrequency items have one possible correct answer out of a total of five response options. The probability of conscientious responses being flagged by infrequency items (a confusion rate) was varied between .00, .05, and .20, to emulate scenarios in which conscientious responders confuse, incorrectly interpret, or idiosyncratically interpret infrequency items 0%, 5%, and 20% of the time.

### Measures

For each cutoff threshold, we calculated several quantities of interest. We calculated the correlation between two substantive variables, the average Mahalanobis distances of the observations removed by each technique, and the classification frequencies of each technique.

<sup>1</sup> A supplementary R package to calculate this quantity can be found at <https://github.com/uwkinglab/detectpme>.

**Correlations** Since the correlation between all variables are the same in the data generating parameters, we computed the correlations between the first two variables without the loss of generality. The correlations are calculated in three ways: (1) using the original sample of conscientious responses only (denoted original), (2) using the combined sample of conscientious and random responses (denoted contaminated), and (3) using the cleaned sample of responses after the treatment of each cleaning technique (denoted cleaned). Additionally, the percentage of correlation recovered is calculated as the difference between the cleaned and contaminated correlations divided by the difference between the conscientious and contaminated correlations. This represents the degree to which the original correlation's magnitude is returned to its original value after cleaning the data.

**Mahalanobis distance** To quantify the degree of randomness of the data screened by each threshold, we used the average Mahalanobis distance (Mahalanobis, 1936). The Mahalanobis distance is a multivariate distance metric, which measures the closeness of an observation's data pattern to the average pattern after accounting for variance and covariance. For example, conscientious responders on average may respond to scales with a "tighter" pattern or consistency. That is, their responses to any given scale will tend to correlate to the degree that the items reflect the same latent construct. Careless responders are presumed to be answering irrespective of item content, thus conversely, their responses are expected to deviate from the scale's measurement structure. In this regard, random responders are assumed to respond more "loosely" (or randomly), which increases their Mahalanobis distances. In terms of indicating random responding, previous research has shown that Mahalanobis distances strongly loads onto an inconsistent-type of random responding in factor analysis studies (Maniaci & Rogge, 2014; Meade & Craig, 2012). In the present analysis, we calculate the average Mahalanobis distances, along with their standard deviations and effect sizes, for the observations screened by each threshold technique.

**Classification frequency** By simulation design, it is known if an observation was generated by the careless or conscientious data generating process. Thus, to measure the classification performance of each threshold technique, we calculated the numbers of observations that were correctly and incorrectly screened by each threshold, to yield confusion matrices. As such, we observe the number of true and false positives as well as the number of true and false negatives.

## Results

The results of this simulation are displayed in Table 1. Using the error-balancing threshold, the cutoffs of one, three, and

five infrequency items are calculated as the leniency needed to accommodate the confusion rates of conscientious responses of 0%, 5%, and 20%, respectively.

The classification frequencies show the degree of accuracy conferred by each method. When the confusion rate of conscientious responses to infrequency items is zero, the point of minimum error yields a cutoff of 1, the same as zero tolerance. Since the cutoff of 1 separates conscientious from random responders with extremely high accuracy, there is no need for a more lenient cutoff. Both the minimum error and zero tolerance provide perfect classification in this sample. However, when the confusion rate increases to .05, we see that the point of minimum error allows for some leniency, providing a cutoff of 3. Comparing this to the zero-tolerance cutoff of 1, we see that the point of minimum error maintains perfect classification, whereas the zero-tolerance technique erroneously removes 88 of the 900 (9.8%) conscientious responses in the sample. This impact is reflected in the correlation, since 89% of the original correlation is recovered using zero tolerance, whereas 100% of the correlation is recovered with the point of minimum error.

Finally, we studied the extreme case, in which conscientious responses had a 20% confusion probability. In this condition, the minimum error threshold erroneously removed 25 of the 900 (2.8%) conscientious responders and erroneously left only 1 of the 100 (1%) random responders. The zero-tolerance method correctly removed all of the random responders, but also erroneously removed 804 of the 900 (89.3%) conscientious responders, leaving only a total of 96 of 1,000 (9.6%) observations remaining in the dataset.

In summary, this simulation study compared the classification behavior of a minimum error threshold and zero tolerance under several circumstances. When the confusion rate is zero, the minimum error threshold will tend to yield the same cutoff as zero tolerance, and both techniques will enjoy very good classification rates. However, as the confusion rate increases, the point of minimum error maintains good classification by adjusting the cutoff, and zero tolerance will increasingly misclassify, depending on how large the confusion rate is. At large confusion rates, the minimum error technique maintains reasonable classification, whereas zero tolerance instead yields unreasonable flagging rates.

## Study 2

### Method

#### *Participants*

We conducted secondary analyses across two independent datasets (Studies 2a and 2b) of college-enrolled young

**Table 1** Study 1 correlations and classification frequencies

Confusion	0.00				0.05				0.20			
Cutoff	EB (1)		ZT (1)		EB (3)		ZT (1)		EB (5)		ZT (1)	
Correlations												
Original	0.25		0.25		0.24		0.24		0.23		0.23	
Contaminated	0.47		0.47		0.42		0.42		0.39		0.39	
Cleaned	0.25		0.25		0.24		0.26		0.25		0.28	
% Recovered	1.00		1.00		1.00		0.89		0.90		0.70	
Distances												
Mean	43.23		43.23		42.52		26.65		41.66		11.80	
SD	10.99		10.99		8.98		19.17		14.81		14.90	
Effect size	3.03		3.03		3.62		0.87		2.14		0.12	
Frequencies												
	Rem	Left	Rem	Left	Rem	Left	Rem	Left	Rem	Left	Rem	Left
Careless	100	0	100	0	100	0	100	0	99	1	100	0
Conscientious	0	900	0	900	0	900	88	812	25	875	804	96

The cutoff row denotes the cutoff technique used, error balancing (EB) or zero tolerance (ZT), and the number of infrequency items corresponding to that threshold. For correlations, original refers to the sample of conscientious responses only, contaminated refers to the combined sample of conscientious and random responses, cleaned refers to the sample of data treated by each cleaning technique and % recovered is the difference between the cleaned and contaminated correlation divided by the difference between the original and contaminated correlation. For the classification frequencies, “rem” refers to the number of observations removed, and “left” refers to the number of observations remaining in the dataset

adults. Participants in both studies were undergraduate students from a university in the Pacific Northwest region of the United States between the ages of 18 and 20. A total of 403 participants from Study 2a (30.8% male) and 727 from Study 2b (48.8% male) were included for analysis. In Study 2a, 51.8% identified as White, 33.6% identified as Asian, and 14.6% reported other or mixed ethnicity. For Study 2b we exclusively recruited Caucasian and Asian participants; 41.3% identified as White, 57.8% identified as Asian or Asian-American, and 1% did not report ethnicity. The participants for each study were recruited via postings on an online subject pool program through the university, and data were collected during single-session Web-based surveys that took approximately 1 h to complete. Students received optional course credit as compensation for their participation. Both studies were approved by the university’s institutional review board.

### Measures

**Infrequency scale** We used the “bogus items” developed by Meade and Craig (2012) as our infrequency scale for both of these studies. The items were spread uniformly throughout each of the surveys. In Study 2a, nine bogus items were implemented with a response set consisting of seven levels of endorsement, from *Strongly Disagree* to *Strongly Agree* (see Table 2). In Study 2b, all ten bogus items were used, with the response set varying to match the response set of the surrounding scales (see Table 3). To code correct and incorrect responses, we maintained

the coding scheme of Meade and Craig (2012), in which only answering with *Strongly Agree* or *Agree* (or their negated and/or most close numerical counterparts) is considered to be correct.

**Ancillary scales** Since the usage of infrequency scales pertains primarily to low-stakes testing, we tested the study hypotheses using three scales typically used in low-stakes testing scenarios, measuring constructs of alcohol problems, parenting, and personality. Two of the scales—the Alcohol Use Disorders Identification Test (AUDIT; Saunders, Aasland, Babor, de la Fuente, & Grant, 1993) and the Parental Bonding Instrument (PBI; Parker, Tupling, & Brown, 1979)—were available in both datasets. The third measure—the UPPS Impulsive Behavior Scale (Whiteside & Lynam, 2001)—was available only in Study 2a. We had no research questions related to these scales or their content.

**Random-responding indicator** To quantify random responding, we used the Mahalanobis distance (Mahalanobis, 1936), as validated by Study 1 of the present article and previous research (Maniaci & Rogge, 2014; Meade & Craig, 2012). However, in Study 2, since the computational complexity of the Mahalanobis distance increases exponentially with the number of variables, we calculated the Mahalanobis distance separately for each scale and then took the average, as in Meade and Craig (2012).

**Table 2** Study 2a infrequency scale frequencies

Item	Responses						
	Strongly Agree	Agree	Slightly Agree	Neither Agree or Disagree	Slightly Disagree	Disagree	Strongly Disagree
1. I am using a computer currently	<b>309</b>	<b>29</b>	6	8	8	6	33
2. I am enrolled in a Psychology course currently	<b>315</b>	<b>44</b>	7	4	6	7	12
	Strongly Disagree	Disagree	Slightly Disagree	Neither Agree or Disagree	Slightly Agree	Agree	Strongly Agree
3. I sleep less than one hour per night	<b>326</b>	<b>34</b>	13	10	7	4	2
4. I do not understand a word of English	<b>365</b>	<b>16</b>	3	6	5	1	3
5. I have never brushed my teeth	<b>358</b>	<b>21</b>	7	7	3	2	2
6. I am paid biweekly by leprechauns	<b>331</b>	<b>22</b>	5	17	5	1	14
7. All my friends are aliens	<b>312</b>	<b>36</b>	14	13	5	5	11
8. I have been to every country in the world	<b>253</b>	<b>95</b>	24	12	11	1	7
9. All my friends say I would make a great poodle	<b>131</b>	<b>57</b>	15	159	11	12	10

\*Boldface indicates a correct response

**Table 3** Study 2b infrequency scale frequencies

Item	Responses								
	Strongly Agree	Agree		Neutral		Disagree		Strongly Disagree	
1. I am using a computer currently	<b>609</b>	<b>62</b>		16		2		9	
	Strongly Agree	Agree	Slightly Agree	Neither Agree or Disagree	Slightly Disagree	Disagree		Strongly Disagree	
2. I am enrolled in a Psychology course currently	<b>616</b>	<b>47</b>	19	15	5	1		3	
	Disagree a Lot	Disagree		Neutral		Agree		Agree a Lot	
3. I have been to every country in the world	<b>504</b>	<b>83</b>		40		25		8	
	1 Never or Definitely No	2	3	4	5	6	7	8	9 Always or Definitely Yes
4. I have never spoken to anyone who was listening	<b>507</b>	<b>37</b>	22	19	27	16	13	7	3
5. I sleep less than one hour per night	<b>616</b>	<b>43</b>	13	9	12	4	8	8	3
6. I do not understand a word of English	<b>634</b>	<b>18</b>	8	5	15	11	6	6	2
	Strongly Disagree	Disagree	Slightly Disagree	Neither Agree nor Disagree	Slightly Agree	Agree		Strongly Agree	
7. I have never brushed my teeth	<b>351</b>	<b>8</b>	5	12	10	3		5	
8. I am paid biweekly by leprechauns	<b>454</b>	<b>14</b>	9	39	11	20		17	
	Rarely or none of the time (less than 1 day)			Some or a little of the time (1–2 days)		Occasionally or a moderate amount of the time (3–4 days)		Most or all of the time (5–7 days)	
9. All my friends are aliens	<b>545</b>			44		20		17	
	Strongly Disagree			Disagree		Agree		Strongly Agree	
10. All my friends say I would make a great poodle	<b>394</b>			<b>90</b>		48		32	

\*Boldface indicates a correct response

### Statistical analysis

We analyzed the response patterns among participants whose data were removed by each screening condition.<sup>2</sup> As we previously mentioned, the outcome variable of the study was the average Mahalanobis distances of each screening condition. We used a randomization test to formally test our hypothesis. A randomization test derives an empirical sampling distribution by calculating a test statistic of interest for all (or a large amount) of the permutations of the data under which the null hypothesis is true (Rodgers, 1999). This is a nonparametric technique that obviates distributional assumptions (e.g., normality), thereby increasing statistical power (Edgington, 1964). This was ideal for the present study, because the distribution of the Mahalanobis distances was skewed, and some of the screening conditions yielded low sample sizes (as low as  $n = 32$ ).

## Results

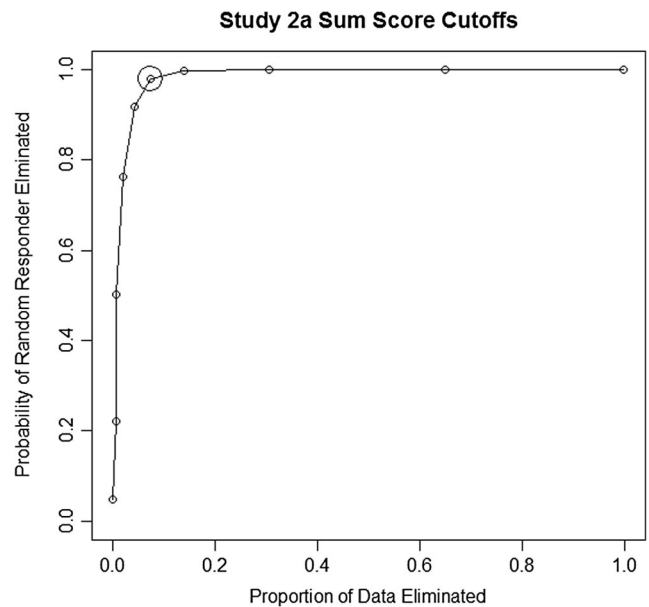
### Threshold calculations

Subsets of randomly responding individuals were identified using the error-balancing threshold and the zero-tolerance threshold. The ROC-like curves for determining the error-balancing threshold are depicted in Figs. 1 and 2. For both studies, the error-balancing optimization procedure established a threshold of four infrequency items answered incorrectly as the cutoff. In Study 2a,  $n = 32$  respondents (8.0%) were identified using this cutoff, and  $n = 52$  respondents (7.2%) in Study 2b. Using the zero-tolerance threshold, in Study 2a,  $n = 272$  respondents (67.5%) were identified, and in Study 2b,  $n = 265$  (36.4%) were identified.

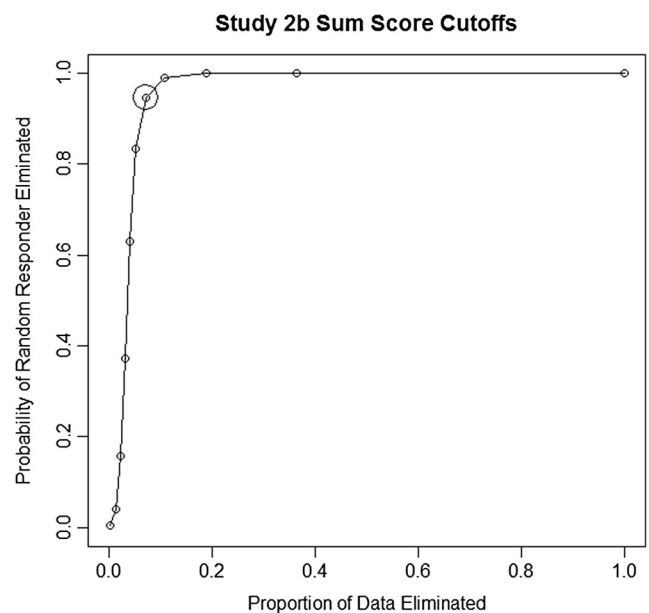
### Mahalanobis distance analysis

The descriptive and inferential statistics of the average Mahalanobis distances in the present study are described in Table 4. In Study 2a, the error-balancing threshold identified respondents with significantly higher average Mahalanobis distances than the null distribution ( $d = 1.215$ ), whereas the zero-tolerance threshold did not ( $d = 0.076$ ). In Study 2b, the Mahalanobis distances were significantly different for both the error-balancing and zero-tolerance thresholds. However, the error-balancing threshold had a much higher effect size than the zero-tolerance condition, consistent with Study 2a ( $d_s = 0.875$  and  $0.257$ , respectively).

<sup>2</sup> Prior to the analyses, the data were examined for long-string response patterns, because of their ability to artificially deflate Mahalanobis distances. The maximum long string for each participant was calculated, and the distributions of the flagged and nonflagged categories were compared. The two distributions were nearly identical to each other, and thus no cleaning actions were deemed necessary.



**Fig. 1** Study 2a sum score receiver operating characteristic (ROC) curve. Each point represents a minimum number of infrequency items answered incorrectly, ranging from 9 to 0, from bottom left to upper right. The point at which the proportion of data eliminated is balanced as being equal to the probability of a random responder being eliminated is 4 or greater. This point is indicated by the surrounding circle



**Fig. 2** Study 2b sum score ROC curve. Each point represents a minimum number of infrequency items answered incorrectly, ranging from 10 to 0, from bottom left to upper right. The point at which the proportion of data eliminated is balanced as being equal to the probability of a random responder being eliminated is 4 or greater. This point is indicated by the surrounding circle



**Table 4** Analysis of average Mahalanobis distances

	Screening	<i>n</i>	%	Observed Mean	Difference From Null Mean	Standard Error	Standardized Difference	Effect Size (Cohen's <i>d</i> )
Study 2a	Error balancing	32	8.0	38.016	13.050	1.909	6.838*	1.215
	Zero tolerance	272	67.5	25.795	0.819	0.603	1.357	0.076
Study 2b	Error balancing	52	7.2	20.415	6.764	1.057	6.397*	0.875
	Zero tolerance	265	36.4	15.645	1.985	0.456	4.357*	0.257

\* $p \leq .001$ 

## Discussion

The present study provides a unique empirical investigation into the usage of infrequency scale thresholds in screening random responders. Our findings indicate that the zero-tolerance cutoff approach was largely outperformed by the error-balancing threshold in both simulated and real data settings. The results from Study 1 showed that the zero-tolerance approach is adversely affected by item confusion rates: At low confusion rates it begins to perform poorly (9.8% false positives), and at high confusion rates it is largely nonfunctional (89.3% false positives). At a confusion rate of zero, however, the thresholds performed identically. We note that this situation may arise in applied settings. Unlike the bogus items currently studied, other infrequency scales may contain items with much more clear answers (e.g., “Please select ‘agree’ for this question”). For such items, it may be the case that the confusion rate is indeed close to zero, and in fact the zero-tolerance threshold may be the appropriate cutoff method. In the conditions of Study 1, we show that the error-balancing threshold correctly reduced to the zero-tolerance threshold. As such, the error-balancing threshold generalizes well to such cases.

In the real data setting of Study 2, the zero-tolerance cutoff identified far more respondents (36%–67% of the samples) than the error-balancing threshold (7%–8%), and generally failed to identify respondents whose scale responses were substantially more random than the average. The large number of samples flagged by the zero-tolerance approach shows a potentially large reduction in power when using this method. These effects are likely driven by the assumptions of zero tolerance, which assumes that infrequency items function perfectly and that answering one infrequency item incorrectly implies all the respondent's survey responses were careless.

It may be the case that invalid responding is more state-like, and can change throughout the course of a survey. For example, it is well known that survey length can detrimentally affect response quality (Galesic & Bosnjak, 2009) and scale validity in real data (Burisch, 1984, 1997). Thus, using responses to an infrequency scale as a binary measure may be unreliable in indicating overall response quality for the entire survey. Indeed, the respondents in the present study who only met the zero-tolerance threshold had low average Mahalanobis distances (meaning their responses were relatively tightly clustered), suggesting that a majority of their responses were actually nonrandom.

The error-balancing threshold performed much better in screening deviant response patterns by relaxing the discard threshold to four items or more. If an infrequency item is causing a threshold to screen a disproportionately large amount of data, then the error-balancing optimization will counter this by allowing more infrequency items to be answered incorrectly. This automatic adjustment is a particular advantage of the error-balancing threshold, in addition to being agnostic to the quality of the infrequency items.

## Recommendations

First and foremost, we recommend against using the zero-tolerance cutoff without justification. Considering the design and logic of infrequency scales, it may seem sensible to use a zero-tolerance threshold, which has become the de facto standard (e.g., Fervaha & Remington, 2013; Osborne & Blanchard, 2011; Periard & Burns, 2014). Our findings suggest that this approach screens too many conscientious respondents, however, relative to methods with more relaxed cutoff points. As infrequency scales gain popularity in the future, we suggest the use of more lenient thresholds in screening for carelessness, and we propose that the error-balancing

threshold provides a good theory-based starting point to determine cutoffs.

### Limitations

**Assumptions** The error-balancing cutoff technique calculates an optimal cutoff under several assumptions. First, it uses the probability of eliminating a uniformly random variable and the proportion of screened data as its performance criteria. Second, it optimizes the cutoff by weighting these to criteria equally. These assumptions were employed using ROC analysis theory, and we believe them to be reasonable approximations. However, as research progresses on the nature of careless responding as a whole, alternate probability models and different weighting schemes should be studied. These can be readily adapted to the error-balancing threshold, or another technique may supersede it altogether. All in all, there has been very little research on optimal thresholds and criteria as a whole, leaving many opportunities for further empirical work.

**Sample generalizability** As with most studies, the utilization of student samples may limit the generalizability of the findings. However, we presume that the present findings should generally apply to low-stakes testing situations, in which responses generally have little external motivation. Although the *rates* of careless responding may vary from sample to sample (Berinsky et al., 2014), we believe it is reasonable that careless responding *indicators* should be comparable, regardless of the sample they were generated from (e.g., a careless responding profile from a student will look similar to one generated by a working adult).

**Study design** As with all observational studies, we note that the results of Study 2 are purely correlational, since it was not known if a participant was in truth conscientious or careless. We further note that only one outcome was studied to indicate careless responding. However, we temper these facts by noting that the results of Study 2 are validated by the simulation results of Study 1, and are consistent with prior literature (Maniaci & Rogge, 2014; Meade & Craig, 2012). Furthermore, we replicated our effects across two independent datasets with differing demographic characteristics.

### Future directions

Several major avenues of research still need to be examined in the usage of infrequency scales. As we previously mentioned, the item functioning characteristics of infrequency items deserves empirical research. An implication of the present study is that infrequency items do not necessarily function as they were intended. As such, using measurement models (e.g.,

structural equation modeling or item response theory) may help relate item functioning to random responding or carelessness as a latent construct.

In terms of careless responding detection methods, only infrequency scales were examined in the present study. The study of optimal screening can easily be generalized toward alternative methods of careless responding detection. For example, instructional manipulation checks (Oppenheimer et al., 2009), psychometric synonyms/antonyms (Maniaci & Rogge, 2014; Meade & Craig, 2012), or even response time analysis (Curran, 2016) may require more thoughtful balancing of classification errors. There has been very little research into how to classify participants on the basis of their responses to these detection methods. As we found in the present study, simply applying a zero-tolerance cutoff method may not be the optimal choice.

Broadly speaking, careless responding screening procedures generally eliminate all data from respondents who are classified to be careless. This approach presumes that all a flagged participant's data are unreliable, implying that the participant was always responding carelessly. However, it is possible that careless responding may operate more as a transient, state-like factor than as a permanent trait. As such, researchers need to understand carelessness as a momentary process or a phenomenon that occurs over the course of the survey (e.g., in response to participant fatigue). The development of more sensitive methods to detect such a model of carelessness is another opportunity for further research.

Finally, in a nascent area of study, the effect of careless responding on statistical estimates is being investigated directly. For example, even at low base rates, careless responding has been found to decrease or spuriously increase both correlations (Credé, 2010; Huang et al., 2015) and confound model selection indices (Woods, 2006). Indeed, the impact of careless responding on the statistical estimation procedure may have unexpected and varied effects, and this is not yet widely understood. One of the goals of detecting careless responding is to improve the quality of statistical estimates, so discovering precise effects in different contexts is an opportunity for future study.

### Conclusion

Infrequency scales are tools that are both convenient and useful for improving data quality. Their utilization, however, is not as straightforward as employing a zero-tolerance threshold. Choosing a proper threshold may seem a trivial issue on the surface; however, the present study has shown that actual indications of random responding may differ greatly. Researchers who wish to utilize infrequency scales should be aware of such issues and the consequences associated with using various thresholds.

## References

- Bakan, P. (1960). Response-tendencies in attempts to generate random binary series. *American Journal of Psychology*, 73, 127–131.
- Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology*, 123, 101–103.
- Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58, 739–753. doi:<https://doi.org/10.1111/ajps.12081>
- Berry, D. T. R., 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, 340–345.
- Brener, N. D., Kann, L., Shanklin, S., Kinchen, S., Eaton, D. K., Hawkins, J., ... Centers for Disease Control and Prevention (CDC). (2013). Methodology of the Youth Risk Behavior Surveillance System—2013. *MMWR Recommendations and Reports*, 62, 1–20.
- Burisch, M. (1984). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality*, 18, 81–98.
- Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality*, 11, 303–315.
- Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, 70, 596–612.
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. doi:<https://doi.org/10.1016/j.jesp.2015.07.006>
- Curran, P. G., & Hauser, K. A. (2015). Understanding responses to check items: A verbal protocol analysis. Paper presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia.
- R Development Core Team. (2015). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org>
- Edgington, E. S. (1964). Randomization tests. *Journal of Psychology*, 57, 445–449.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874.
- Fervaha, G., & Remington, G. (2013). Invalid responding in questionnaire-based research: Implications for the study of schizotypy. *Psychological Assessment*, 25, 1355–1360. doi:<https://doi.org/10.1037/a0033520>
- Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73, 349–360.
- Hathaway, S., & McKinley, J. (1951). *Minnesota Multiphasic Personality Inventory; manual (Revised)*.
- Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100, 828–845. doi:<https://doi.org/10.1037/a0038510>
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39, 103–129.
- Johnston, L. D., O'Malley, P. M., Miech, R. A., Bachman, J. G., & Schulenberg, J. E. (2016). Monitoring the future national survey results on drug use, 1975–2015: Overview, key findings on adolescent drug use (Report?). Ann Arbor: University of Michigan, Institute for Social Research.
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences*, 2, 49–55.
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61–83.
- Marjanovic, Z., Struthers, C. W., Cribbie, R., & Greenglass, E. R. (2014). The Conscientious Responders Scale: A new tool for discriminating between conscientious and random responders. *SAGE Open*, 4(3). doi:<https://doi.org/10.1177/2158244014545964>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17, 437–455.
- Nichols, D. S., Greene, R. L., & Schmolck, P. (1986). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, 45, 239–250.
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45, 867–872.
- Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*, 1, 220:1–7. doi:<https://doi.org/10.3389/fpsyg.2010.00220>
- Parker, G., Tupling, H., & Brown, L. B. (1979). A parental bonding instrument. *British Journal of Medical Psychology*, 52, 1–10. doi:<https://doi.org/10.1111/j.2044-8341.1979.tb02487.x>
- Periard, D. A., & Burns, G. N. (2014). The relative importance of Big Five Facets in the prediction of emotional exhaustion. *Personality and Individual Differences*, 63, 1–5.
- Robinson-Cimpian, J. P. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educational Researcher*, 43, 171–185.
- Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, 34, 441–456.
- Saunders, J. B., Aasland, O. G., Babor, T. F., de la Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction*, 88, 791–804.
- Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, 16, 73–81.
- Tune, G. S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin*, 61, 286–302.
- Wang, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica*, 3, 295–312.
- Whiteside, S. P., & Lynam, D. R. (2001). The Five Factor Model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, 30, 669–689. doi:[https://doi.org/10.1016/S0191-8869\(00\)00064-7](https://doi.org/10.1016/S0191-8869(00)00064-7)
- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28, 189–194. doi:<https://doi.org/10.1007/s10862-005-9004-7>
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3, 32–35.