Journal of Research in Personality 82 (2019) 103849

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Full Length Article

I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items

Paul G. Curran [a,*], Kelsey A. Hauser [b]

[a] Grand Valley State University, Allendale, MI, United States
[b] George Washington University, Washington, DC, United States

ARTICLE INFO

Article history:
Received 25 September 2018
Revised 14 June 2019
Accepted 25 July 2019
Available online 26 July 2019

Keywords:
Carelessness
Data cleaning
Insufficient effort responding
Verbal protocol
Self-report data

ABSTRACT

Participant carelessness is a source of invalidity in psychological data (Huang, Liu, & Bowling, 2015), and many methods have been created to screen for this carelessness (Curran, 2016; Johnson, 2005). These include items that researchers presume thoughtful individuals will answer in a given way (e.g., disagreement with "I am paid biweekly by leprechauns", Meade & Craig, 2012). This paper reports on two samples in which individuals spoke aloud a series of these questions, and found that (a) individuals do occasionally report valid justifications for presumed invalid responses, (b) there is relatively high variance in this behavior over different items, and (c) items developed for this specific purpose tend to work better than those drawn from other sources or created ad-hoc.

1. Introduction

It is impossible to argue that self-report psychological data always perfectly reflects the underlying psychological constructs which it is trying to measure. There are many reasons for this disconnect, and many of these have been well-studied for decades (e.g. Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989; Dunnette, McCartney, Carlson, & Kirchner, 1962; Orpen, 1971). Not surprisingly, the vast majority of this longstanding research focuses on situations where questions are being asked and answered in high-stakes situations, such as job applications (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006) or clinical assessment (Berry, Baer, & Harris, 1991). In these situations of faking, survey-takers may actually be putting in more effort to answer disingenuously than would be required to answer truthfully.

The opposite end of this spectrum occurs when survey-takers put in less effort than is necessary to answer truthfully or thoughtfully (Huang, Curran, Keeney, Poposki, & DeShon, 2012, Huang, Liu, & Bowling, 2015, Meade & Craig, 2012). Research on detecting and/or deterring these particular invalid responders has been on the rise in recent years, and several terms for these individuals have entered common parlance. These include 'careless responder' (Meade & Craig, 2012) and 'insufficient effort responder' (Huang et al., 2012), or a combination, "C/IE responder" (Curran, 2016), which we will use here.

There are many proposed methods of detecting C/IE responders (Curran, 2016; Johnson, 2005), and one popular method is the inclusion of extra items in the scale meant to act as alarms. These items range from "Please select 'Moderately Inaccurate' for this item" (Huang et al., 2012), to "I am interested in pursuing a degree in parabanjology" (Huang et al., 2014), to "I am paid biweekly by leprechauns" (Meade & Craig, 2012).

Of the many other techniques that can be used to detect C/IE response (e.g., psychometric antonyms, odd-even consistency, Mahalanobis distance), placing questions of this sort in a survey is relatively transparent. Unlike an item assessing an individual difference (e.g., 'I am the life of the party'), these types of items have responses which are considered to be more correct than others. In the case of items with an instructed response (e.g., 'Please select. . .'), there is one correct answer that it is assumed a thoughtful responder would provide. In the case of other items without instructed responses (e.g., 'I am currently using a computer.'), thoughtful respondents are assumed to uniformly agree or disagree with the statement (depending on the content).

Despite the content of each item, they are all striving toward the same goal: provide survey-takers a place to get an item 'wrong' as a means of detecting lapses in attention. Someone who disagrees to an item about their computer use (as above) while taking the survey on a computer is assumed to be responding without reading the item. We posit, however, that this is not the only

* Corresponding author.
E-mail addresses: paulcurranatwork@gmail.com (P.G. Curran), kahauser@gwu.edu (K.A. Hauser).

https://doi.org/10.1016/j.jrp.2019.103849
0092-6566/

way that respondents to these questions could arrive at their 'incorrect' responses. Using one of these items to screen for attention does not allow for a distinction between inattention, over-thinking, or even temporary rebelliousness. This problem begs the question of what survey-takers are thinking when they chose a particular response, and if some of these items are more prone to these problems than others.

It is possible that these items do not have a significant problem of this sort, that this type of behavior is rare enough only to be found in large or quirky samples, and can be seen as negligible to study outcomes. However, if some thoughtful individuals do respond to these attention check items in a way that is scored incorrect, thoughtful responders may be removed from the sample inadvertently; throwing out the baby with the bathwater, if you will. These are false positives of the data cleaning process, and these are the participants that are the prime focus of this study.

The presence of such behavior at non-negligible rates in normal samples would have strong implications for the use of these items as a data-cleaning tool, as at the moment there is no understanding in the literature of what false positive rates these items might produce. The best guess of the authors of this paper is that there is a broad assumption that these items have a false positive rate at or near zero, as researchers assume that no one who is acting thoughtfully, and non-maliciously, would ever, for instance, agree to an item such as 'all my friends are aliens' or 'I am paid biweekly by leprechauns.'

The impact of non-negligible false positive rates here would drive new work in careless responding detection, as it would point to a potential flaw that has been so far unexamined. At the same time, there would potentially be a broader impact on research studies that use these types of items for data cleaning, particularly in small quantities and without using other careless detection techniques. Using an item with a false positive rate of even 5% would mean that either (a) any sample collected needs to be 5% larger to achieve the same power, or (b) that any sample cleaned with that item is going to be 5% smaller (and that the statistical power of that study will drop accordingly). That impact of an item with a 5% false positive rate is also no means the floor of this effect – the potential impact could be even higher.

The deeper problem here is that we simply do not currently know if this is a problem, and if so, what the magnitude of that problem might be. A 5% reduction in sample size may not be enough to significantly impact the power of any given study, but a 10%, or 15%, or even greater reduction likely will be. The impact of this paper comes in part from highlighting and attempting to address this question, even if the answer to this question is that there isn't much of an impact.

Despite concerns from the authors of these types of items about the interchangeable nature of different items and the thought process survey-takers might use to respond to them (Huang et al., 2014, Meade & Craig, 2012), no work has yet simply asked survey-takers this question to see what they are in fact thinking when they respond to these items, or compared responses between parts of these different scales.

We believe the next step toward understanding of these types of survey items as tools for detecting C/IE responders is the use of protocol analysis (Sudman, Bradburn, & Schwarz, 1996), or similar heuristics. This content analysis of survey-takers' actual thoughts provides the opportunity to elicit from respondents the simple answer to the question: 'what were you thinking when you chose *that* response?'

The answer to this question, at the participant and item level, allows for the comparison between two different sources of information regarding the validity of that participant's response: (1) whether or not the response is 'correct' for that item, and (2) whether or not the participant provided a valid reason for choosing that response, for that item. The first is the information that researchers have when using these items in research and practice, and the second can be considered a proxy for the actual validity of response.

Two of the main ways that C/IE metrics are evaluated are through the calculation of sensitivity and specificity of those metrics, or rather through the true positive rate (sensitivity) and true negative rate (specificity). True positives are when measures of C/IE actually detect participants who are responding carelessly. True negatives are when measures of C/IE fail to retain participants who are responding thoughtfully.

Both of these rates can be assessed in C/IE research by generating (either through simulation or instructing participants to respond carelessly) careless data that is then mixed with some source of 'good' data. From this, the rate that these measures can detect this 'bad' data or retain 'good' data can be calculated (e.g., Meade & Craig, 2012). Doing nothing regarding C/IE responders produces a true positive rate of 0 and a true negative rate of 1; doing anything else begins to (hopefully) increase the true positive rate and (potentially) decrease the true negative rate.

This paper has little interest in evaluating the true positive rate of these types of items. This question is already implicit in the initial studies of these types of items (Huang et al., 2014, Meade & Craig, 2012). What has not been assessed in these studies is what proportion of thoughtful individuals may be incorrectly identified as careless by these items (i.e., false positives; Type I errors). Put another way, what proportion of individuals who are responding thoughtfully may actually choose and justify a response that would be considered 'incorrect' by a researcher using these items?

It is the analysis of this justification of response that is novel to the study of items of this nature and what this study stands to contribute to this literature. A number of outcomes are possible given participants' justifications to their selected responses. Participants may answer with an expected response (e.g. 'strongly disagree' to "I am paid biweekly by leprechauns"), and also provide justification for that response (e.g. 'I'm choosing strongly disagree to this item because leprechauns don't exist'). This can be viewed as a true negative, as the item would classify this participant as a thoughtful responder, and their response showed this thoughtfulness.

A true positive is the direct opposite, a situation where both response choice and verbal justification identify the participant as a C/IE responder. It is more difficult to contrive an example of what this justification would look like, largely because *any* justification is an ostensibly thoughtful act, and thus precludes that justification from indicating carelessness. This is a blind spot of this study that we recognize and for which we do not have a clear solution. This blind spot also obscures false negatives, situations where individuals' justification classifies them as careless while their response does not.[1]

Instead, the focus of this paper is the evaluation and identification of false positives, where a participant provides what would normally be considered an invalid response that they then adequately justify. The goal of this paper is to examine participants' motives behind these 'incorrect' responses, and show that normal, thoughtful participants may be generating these 'incorrect' responses more than we might assume. If this is the case, further studies examining the properties of these items are likely warranted, and use of these items is likely to need stronger guidance and guidelines.

[1] This is not to say that this study is completely blind to this, as participants could provide poor justification or inconsistent justification. Rather, we simply acknowledge that these rates would not be reliable estimates given the nature of our sample providing justification of their responses, and also that the true positive rate is not the focus of this paper – the false positive rate is.

While there is no firm assumption in the literature regarding the false positive rate of these items, part of that absence of assumption is perhaps because of the sheer transparency of these items. That is, if someone doesn't respond with 'Agree' to an item telling them to do so, we assume they are responding carelessly. In this way, we may in fact be assuming that the Type I error rate of these items is 0 – if someone reads the item we assume they will do what they are instructed to do.

Because of this, the statistical goal of this paper becomes one simply of an existence proof. Our main research question can simply be framed as: 'does this happen?' Thus, we propose that there exist, in reasonably-sized, normal samples (e.g., college students), individuals who will be flagged as careless by an attention check item for a response that that participant can adequately justify.

Regarding sample size, our samples are smaller than might be expected in a study of this type examining other qualities of these items (e.g., psychometric properties of items; factor analysis). It would be easy to overpower this research question by collecting many subjects with the goal of looking for a needle in a haystack. We are not attempting to simply show this behavior exists at a population level, or to definitively show the base rate, but rather to demonstrate that this is potentially an issue for anyone using these types of items, even if their sample is of a modest size. Because of this, our sample size should not be terrifically large, only large enough to demonstrate this existence in what could be considered a possible sample for a modest but acceptably small study.[2] This couples well with how our data is collected (verbal protocol analysis), as this method of data collection meant that participants could only be assessed one at a time, in person, and that each participant's recorded audio was coded by the researchers. Weighing these concepts, we decided a priori that a sample size of 60 would be adequate to address this existence proof, of which we were only one participant shy.

In choosing items for this study we also wanted to evaluate how well some of the commonly used items in C/IE research would fare against other similar items from other established scales, as well as from ad hoc, researcher generated items. Researchers who are not familiar with this literature may still independently derive this technique, creating their own items. Creating our own ad hoc items was meant to mimic this process.

This first group of items contained items from Huang et al. (2014) and Meade and Craig (2012), as well as standard 'Answer with…' style items (also known as 'instructed response items'). In addition, we pulled items from other similar scales (see Method), as well as generated a number of items that a researcher might craft on their own in the spirit of the above items.[3] The items that are published in this literature should presumably be those that have already made it through some level of validation, and as such we would suspect that they should perform better than these other items. This leads to a number of secondary research questions:

Is there reasonable variance between items regarding their Type I error rates?

Is the Type I error rate for established items lower than for other items?

2. Method

Two distinct samples were collected for this study, and data from the first was examined before making decisions about the nature of the second. Both samples address the same research questions, and will therefore simply identified as sample 1 and sample 2.

2.1. Participants

Sample 1. Participants consisted of 59 students from a small Midwestern liberal arts college. Participants were largely female, white, and college aged (18–22). All participants received psychology research credit for their participation. Participants were, to the best of our knowledge, no different than those who might be used for any other study of C/IE detection, or for any other psychological research.

Sample 2. Sample size goals were set in sample 1 without knowledge of what these rates may look like in actual sample data. For sample 2, these base rates were able to be estimated from sample 1. Because of this, and because sample 2 was largely designed to simply back up the results of sample 1, it was decided that a sample size of 30 participants would be a reasonable goal. This sample size goal was met, and data collection was terminated at that size.

Participants consisted of 30 students from a small Midwestern liberal arts college. These students were all distinct from those who participated in sample 1. Participants were no different from sample 1 in demographic data, and all received psychology research credit for their participation. Again, participants were no different than those that might participate in any given psychological study.

2.2. Measures

Sample 1. To examine these research questions, we constructed a 40 item battery consisting of items from a number of different sources, including items that are commonly used for C/IE detection and those that could reasonably be drawn from other sources. The final list of these items can be found in table 1, and consists of items from the following sources:

1) Eight items from Huang et al. (2014). An example item is "I work fourteen months in a year."
2) Nine of the ten items from Meade and Craig (2012). An example item is "All my friends are aliens."
3) Five items of the form: "Answer with 'agree' for this item."
4) Four items from the Unusual Beliefs scale of the CAT-PD (Simms et al., 2011). An example item is "I am able to read the minds of others."
5) Three items based on Hargittai (2009) knowledge faking scale. An example item is "I am familiar with geological terms such as jpeg and firewall."
6) Eleven items generated by the authors in keeping with the general framework of items that may be used for C/IE detection. An example item is "Fish live in water."

All items were administered in a random order to each participant using Qualtrics. Response options were on a Likert scale from 1 (strongly disagree) to 7 (strongly agree).

Sample 2. In addition to those items from sample 1, participants were also administered 100 items from the International Personality Item Pool (IPIP) designed to measure the big five factor structure (Goldberg, 1999; Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006). This included 20 items measuring each of the big five factors (conscientiousness, agreeableness, neuroticism, openness to experience, and extraversion). There are no hypotheses relating to standings on these items; they were simply included to extend the generalizability of sample 1.

[2] For instance, a sample size of 60 would give 0.80 power to a study looking for a correlation of approximately $r = 0.30$, or a Cohen's d of approximately $d = 0.70$.

[3] Anecdotally, the authors have encountered a number of colleagues at a range of institutions that use these types of questions for data cleaning, but do not use a standard set, rather using items that they have generated on their own (e.g., 'A hat is worn on your head.').

Table 1
Breakdown of outcomes by item, by scale.

	Sample 1		Sample 2
Item	False Positive Rate	False Negative Rate	False Positive Rate
Huang et al. (2014)			
I work fourteen months in a year	0%	3%	3%
I have never used a computer	0%	5%	0%
I work twenty-eight hours in a typical work day	3%	3%	0%
I can run 2 miles in 2 min	3%	3%	0%
I can teleport across time and space	3%	4%	13%
I eat cement occasionally	7%	3%	7%
I will be punished for meeting the requirements of my job	10%	7%	13%
I am interested in pursuing a degree in parabanjology	41%	7%	27%
Meade and Craig (2012)			
I have been to every country in the world	0%	2%	0%
I sleep less than one hour per night	2%	3%	3%
I am paid biweekly by leprechauns	2%	5%	3%
I have never brushed my teeth	0%	7%	0%
I am using a computer currently	0%	7%	3%
I do not understand a word of English	3%	5%	7%
I have never spoken to anyone who was listening	7%	7%	7%
All my friends are aliens	7%	8%	14%
All my friends say I would make a great poodle	14%	3%	28%
Instructed response items			
Answer with 'Agree' for this item	2%	37%	7%
Answer with 'Disagree' for this item	3%	36%	7%
Answer with 'Strongly disagree' for this item	3%	36%	10%
Answer with 'Strongly agree' for this item	8%	32%	17%
Answer with 'Neither agree nor disagree' for this item	5%	36%	3%
"Unusual Beliefs" Scale			
I have the power to cast spells on others	3%	5%	7%
I believe I have supernatural powers	5%	5%	21%
I can see into the future	10%	5%	23%
I am able to read the minds of others	12%	5%	10%
Hargittai (2009) type items			
I am familiar with computer terms such as filtibly and proxypod	5%	7%	10%
I am familiar with geological terms such as jpg and firewall	53%	5%	48%
I am fluent in combinatorial English	85%	2%	52%
Items generated by the authors			
I get almost no sleep every night	14%	5%	13%
I can run a mile in 4 min	24%	5%	7%
I can eat as much as a horse	42%	5%	24%
I own a pet stingray	0%	7%	10%
I frequently fear that cows are going to take over the world	2%	5%	3%
I am participating in an online study currently	2%	7%	3%
A dolphin is an animal	0%	8%	3%
Trees are a source of wood	2%	8%	10%
Fish live in water	2%	8%	3%
Humans eat food	2%	8%	0%
Oranges are fruit	0%	10%	3%

2.3. Procedure

Sample 1. Participants signed up for this study online, and scheduled a time to come into the lab for the remainder of the study. Upon arriving at the lab, participants were seated at a computer and given instructions to complete the above questionnaire while speaking both the questions and their justification for response aloud. Participants were informed that their spoken responses would be recorded for later coding. Participants were given an opportunity to ask any clarifying questions, then left alone in a room with the computer to complete the survey and task. Participants were debriefed as to the purpose of the study at the end of data collection, and given an additional opportunity to delete their recording if they so chose (none did).

Following this, participant responses were scored according to whether or not they were the expected 'correct' responses for each item. For items where multiple responses on one side of the scale could be considered correct, we scored these items consistent with Meade and Craig (2012). That is, if the Agree end of the scale was considered correct, then Strongly Agree and Agree were scored as thoughtful response, and the remainder of options were scored as an 'incorrect' response.

For items with only one correct response (that is, items of the form "Answer with 'agree' for this item."), these items were scored as to whether the individual selected that response (correct) or any other response (incorrect).

In addition, audio recordings of spoken responses were coded by the authors regarding whether or not participants justified the response that they selected in some way. In essence, these responses were being coded as to whether or not participants supplied any justification to their selection of an answer. Any rater disagreement was settled through consensus discussion.

Sample 2. One of the potential limitations of sample 1 was that items were administered in a way that would be unlikely to occur in practical use. Instead, these items are likely to be encountered embedded in self-report personality items measuring personality constructs such as conscientiousness or agreeableness. It is possible that participants treated these items in a novel way due to the way they were administered in sample 1.

To examine this question, sample 2 was designed as a follow up with items administered in a way much more consistent with

normal self-report survey practices. That is, items were embedded in a larger set of items measuring the personality constructs of the big five. The goal of sample 2 was to examine whether the general pattern of responses found in sample 1 would hold in a slightly more natural setting. Participants still spoke aloud their responses and justification, so this sample is still likely to be more motivated than usual.

The procedure for sample 2 was identical to that of sample 1, as the only change was to the number of items seen by participants. Instead of responding to 40 items, as in sample 1, participants in sample 2 spoke aloud their responses to 140 items – 100 measuring personality in addition to the 40 from sample 1.

In sample 1, each singular response was coded regarding the justification given by participants, as it was unclear if participants would justify their responses to all items. With the exception of the instructed response items (e.g., "Answer with 'agree' to this item."), participants overwhelmingly justified their responses to all other items. As such, participants' recordings were coded in study 2 only in aggregate. That is, participants' recordings were examined to ensure that no one was responding carelessly, overall. Of our 30 participants, none were. In fact, one participant spent so much time justifying their responses that we had to stop them for reaching the end of their allotted session after only seeing half of the items. Their data were retained for those items they encountered.

3. Results

3.1. Sample 1

The main research question of this study was if these types of items will sometimes identify responses as C/IE when in fact they are actually valid and thoughtful. In aggregate, the answer to this question is yes. While there were a small number of items with no examples of Type I error, the vast majority of items had at least a few individuals providing valid justification for 'incorrect' responses. Full breakdown of results across all items can be found in Table 1.

In general, our secondary research questions are also answered by these data. We expected to see variance over items in Type I error rates, and this is supported. This variance is also not just between scales, but sometimes even between those items in a singular established scale (e.g., 0% to 41% in the items of Huang et al., 2014). There are a number of items across these scales that have a 0% false positive rate, indicating that misidentification of respondents with these items should be rare (on average, lower than around 1 in 60). However, there are also many items with non-zero false positive rates.

Our research question regarding established items vs others also tends to follow our expectations. In general, items from the two scales commonly used for this purpose (Huang et al., 2014; Meade & Craig, 2012) perform well. Notable exceptions are the items "I am interested in pursuing a degree in parabanjology." (41% false positive rate) and "All my friends say I would make a great poodle." (14% false positive rate).

Items of the form 'Answer with…' performed well in terms of false positive rates, though had much higher false negative rates than any other items. This represents a situation where participants chose the correct response, but did not provide justification. This was unexpected, and marks a bit of an odd break between these items and those from the established scales above.

Discounting this false negative rate, data would suggest that items borrowed from other areas and those generated by the authors did not in general perform as well as items generated for this purpose, though there was not as clear of a break between these scales as we would have expected.

From those items borrowed from other scales, the best item was "I have the power to cast spells on others" from the unusual beliefs scale, with a 3% false positive rate. The worst item was "I am fluent in combinatorial English" with an 85% false positive rate.

From those items generated by the authors, the best items were a number with a 0% false positive rate ("I own a pet stingray", "A dolphin is an animal", "Oranges are fruit"), while the worst was "I can eat as much as a horse" with a 42% false positive rate.

In all cases, the false positive rate was driven by valid justification for what would be considered an incorrect answer to that question. Selected examples of some of these justifications to specific items can be found in Table 2. These were not the only justifications for these responses, but are put forth as a representative set to help understand some of the 'why' of participants responses. We believe the narrative that these representative responses starts to paint is not one of isolated cases or excessive creativity, but rather the type of deliberative thoughtfulness that we would actually hope for in our 'good' participants. Interestingly, then, these individuals who are simply being diligent and thoughtful to all items are potentially those at most risk to be incorrectly identified as a false positive careless responder in this fashion.

3.2. Sample 2

In general, the findings of sample 2 mirror those of sample 1. This behavior was detectable even in a small sample of only 30 individuals. Scales created for this specific purpose (e.g., Huang et al., 2014; Meade & Craig, 2012) tended to work well, with again some items as exceptions ("I am interested in pursuing a degree in parabanjology." and "All my friends say I would make a great poodle" were again the weak points of each scale). Five items across the two of these scales did not incorrectly identify a single individual at this sample size, and an additional four items only identified one individual. There was only one other item on the remaining scales ("Humans eat food") that succeeded in failing to incorrectly identify any individuals.

The items generated by the authors again performed moderately well, with the repeat exception of the item "I can eat as much as a horse." As a group, these items were slightly weaker than the two scales above, with a mix of items performing from quite well to somewhat poorly.

The items created by the authors in the style of Hargittai (2009) performed worse than any others, with two of the three incorrectly identifying nearly half of the participants as careless. It is worth noting that this is not to indicate that Hargiattai's original items used for their purposes are flawed, but perhaps more so that it is tricky to generate new items for these uses, even with items to use as a template.

Items from the unusual beliefs scale also incorrectly identified a number of individuals, ranging from two on one item to seven on another. Participants tended to take a less literal stance on these items than is likely intended for their use, perhaps due to their embedding into other items that require a non-literal stance (e.g., "I am the life of the party."). This may have implications for how items of this type are used in the detection of carelessness.

Finally, and again oddly, are the instructed response items. While these items were not coded individually as in the first study, it was the case that participants again had a pattern of neglecting justification for these items. Beyond that, none of these items failed to incorrectly identify any individuals, as several other items did. When designing these studies, it is very likely that we would have identified these items as the most 'safe' in terms of detection. However, each identified at least one individual as careless, despite these individuals reading and responding to 135 other items in a thoughtful way.

Table 2
Selected examples of valid justifications for 'incorrect' answers.

"All my friends are aliens"
'Aliens' is a relative term; I don't actually know for sure
What does that even mean, we're all aliens if there's other life out there
"I am interested in. . .parabanjology"
Might be real so don't want to disagree
It sounds like it could be interesting
"I work twenty-eight hours in a typical work day."
It feels like that sometimes
"I am familiar with geological terms such as jpg and firewall."
I know what those are, but don't know that they're geological
"I am fluent in combinatorial English"
I'm fluent in English
"I am able to read the minds of others"/ "I can see into the future"
Understand general idea of what others are thinking
Close friends know each other
Can plan and expect future events
"I sleep less than one hour per night"
When I'm pulling an all-nighter I do
I sleep very few hours each night
"All my friends say I would make a great poodle"
They say I'm like a puppy
They say I'd make a great koala
Friends say I share dog-like personality
Friends have said my hair looks like a poodle
Have been told I'd make a good dog
Don't know, I've never asked them
"I eat cement occasionally"
There was cement in my braces, sure that I ate some
There are a lot of things that are in cement in a lot of foods, so maybe eating parts of it
"Answer with 'Disagree' for this item"
Item doesn't say how much to disagree (picked 'Strongly disagree')
"I am paid biweekly by leprechauns"
I am paid biweekly, just not by leprechauns
"I can run 2 miles in 2 min"
It doesn't say run with your feet, can do it in my mind
"I have been to every country in the world"
I've been to a lot of countries
I have probably been to more countries than most people
"I can teleport across time and space"
Well, time passes, and I can move places, so that's sort of true
Is walking a type of teleportation?
In my dreams I can because one of my life goals is to be the doctor's companion

4. Discussion

The main goal of this paper was to examine whether or not these types of items have the potential to misclassify individuals who provide valid justifications to otherwise strange responses. Our data show that this is a risk, even in relatively small samples. It is also the case that this rate varies quite substantially by item, with some items showing no risk, and some showing quite high risk. This is an answer to one of our secondary research questions regarding this variability.

Secondary research questions also examined how variable this error rate was over different types of items, specifically those created for this purpose vs those created ad hoc or taken from other similar literature. Established items tended to work better in general, though the effect was perhaps not as large as the authors would have initially expected. For instance, the items generated by the authors in the style of these scales did almost as well as a group as these other two scales. Each of these three scales had at least one item that performed worse than the rest, potentially identifying almost half of the sample in two cases, showing that these scales are not without their faults. In addition, the fact that the author-generated items performed slightly worse than those established items suggests that even individuals entrenched in this research are not capable of writing these items on the fly without needing to validate them in some way.

What do these results mean for the use of these types of items? The results of this study should not be taken to say that these types of items should never be used. In fact, quite the opposite is true. The authors stand by the use of these types of items for the detection of carelessness in self-report survey data. However, these items should not be used without understanding of their limitations and potential errors that can result from misuse. If these items are to be used to screen individuals out of a dataset, we have an obligation to examine the types of errors that can be made.

While the authors will stand by the statement that these types of items can safely be used, there are some potentially more nuanced points that can be taken from our data. In general, items created for this task and validated in published work are likely to be the safest bet for researchers looking to screen data with these types of items, though users of these items should be careful even when using these scales, as there are items that are weaker or stronger than others. Excluding one or two items from each of these scales has the potential to reduce this type of error, as long as you are removing those items that are actually weaker. A mix of the strongest items from multiple scales has the potential to act as a strong way to detect carelessness without detecting unexpected but valid responses.

A careful reader may be at this point asking the question 'doesn't the use of these items as scales protect against this type of error by assuming these types of errors are non-correlated

within thoughtful individuals?' Put simply, yes. Put more complexly, it's complicated.

While any one item may have a reasonable chance of misclassifying an individual, a set of items has a greatly reduced chance[4]. Instead of one item that only has a yes or no opportunity of incorrect detection, a scale of these items produces a score that is not necessarily all or none. Missing one out of five items of this sort is often viewed as forgivable, which seems to match with our results.

To use the nine items of Meade and Craig (2012) from study 2 as an example, there are no participants who missed all nine items, and there are no participants who are even close to such a feat. The cut score for a scale like this would never require a participant to miss all items to be removed, as even computer generated random responses should pass a few of these by chance. While a reasonably conservative cut score of four or five wrong out of nine would protect against this type of error almost completely, a lower cut score of two or three might not. Removing individuals who missed only one of these items on the scale would generate error rates not just in line with our estimates, but much higher, as the error rate would be additive upon uniqueness. Thus, the issue moves from not just the items themselves, but how cut scores are generated and used.

Using nine of these items from a published scale is likely among one of the best case scenarios, but it is also possible for researchers to use only one or two items (for space or time reasons) from a scale, or create one or two items on their own. These are the types of situations that this paper is meant to caution against. Using a single item to detect carelessness will, in even reasonably sized samples, identify some individuals as careless who are providing valid responses. The magnitude of this error will be directly related to the quality of the item that has been picked for this purpose. If an item for this purpose is picked at random, or created by a researcher without proper consideration of how a participant may use it, the potential impact could be quite large. This impact may also lead to different types of behaviors. An item with only a relatively small amount of error (<10%) may simply be seen to indicate that this sample is particularly careless (as this error would be added to those careless individuals being removed for actually missing that item). If the error was considerably higher (>30%), as a number of these items are, the researcher may simply end up discarding the item, along with their hope of excluding those truly careless individuals. This study provides an estimate of this error for those items we examined, but using items that have not been examined in this way produces an unknown error rate that should be examined by any researcher looking to use small groups of their own items. Put simply, a point that we would like to emphasize is to know your items. As an extension of this, know the risk of not knowing your items.

While this paper could never fully examine the set of possible items that could be written for this purpose, it is worth pointing out what we can learn from the set that was examined. In particular, a reader of this paper may be doing so looking for a small number of items that they can use as a screen in their own research. With that in mind as a possible takeaway from this paper, the items with the least error of this type can be found in Table 3. These items all had fewer than 5% (summed over samples) of false positives, representing the lowest rates of any items. Some of these items even had a 0% rate, meaning that no individual in either of our samples gave reasonable justification to selecting an answer that would be scored as careless.

While the authors would still caution against using any single item, a small subset of these items from Table 3 could be used to screen for inattentiveness in a study with minimal risk of this type of error. Using 3–5 of these items and allowing for a reasonable cut score would likely reduce this type of error to infinitesimal levels.

Table 3
Items with fewer than 5% false positives summed over study 1 and study 2.

	Sample 1	Sample 2	Sum
Huang et al. (2014)			
I work fourteen months in a year.	0%	3%	3%
I have never used a computer	0%	0%	0%
I work twenty-eight hours in a typical work day	3%	0%	3%
I can run 2 miles in 2 min	3%	0%	3%
Meade and Craig (2012)			
I have been to every country in the world	0%	0%	0%
I have never brushed my teeth	0%	0%	0%
I am using a computer currently	0%	3%	3%
Author-generated items			
A dolphin is an animal	0%	3%	3%
Humans eat food	2%	0%	2%
Oranges are fruit	0%	3%	3%

The items from Table 3 can also be examined to attempt to discover what makes some of these items less vulnerable to this type of error than others. While this is not a main goal of this paper, it is again something that can be examined in this pool of items. That said, this examination is largely subjective, and a thoughtful reader may come to completely different but equally valid (or more valid!) conclusions regarding the content of these items.

In general, our take on these items is that they all seem to share a common theme not only of impossibility/truth, but of simplistic impossibility/truth. That is, while two statements may both be impossible, some could potentially be seen as having more wiggle room than others. The statements of 'humans eat food', or 'oranges are fruit', 'a dolphin is an animal' have no loopholes, they're simply linguistically true.

The same is true of items that talk about literal impossibilities, such as '. . .fourteen months in a year' or '. . .twenty-eight hours in a day.' There's a little more room in these statements to read something like '*I feel like I* work fourteen months in a year', but at the core they are still breaking rules of how we track time in very simple ways.

Two of these items are also very similar items from two different scales, each asking about the use of a computer. It should be noted that these items are likely less useful with every passing day, as more and more online surveys are completed on mobile devices. In this study, participants were brought to a lab and used a computer, but other studies may not always have that capability. That said, these two statements again represent simple facts without much room for overthinking things.

The three remaining items ('I can run 2 miles in 2 min', 'I have been to every country in the world', and 'I have never brushed my teeth') are somewhat distinct in that while they are likely impossibilities in this sample, they do have the capacity to be individual-dependent. Granted, if someone had achieved the first of these, we would know about it in the record books,[5] but at the same time an individual who was quite fast might not completely disagree with this item. There are likely individuals on the planet (but not in our sample) who have gone to every country in the world, or who have never brushed their teeth. Thus, in more diverse samples these questions may actually start to weaken more than some of the others in this list. It is also the case that an individual may have been to many countries of the world, just not all of them. One a seven-point scale

[4] Results showed very little evidence of persistent false positives by the same participants. That is, these results are not likely due to simple creativity from a few participants who had fun with these items.

[5] For reference, as of this writing, the current world record pace for a mile race is still above three and a half minutes.

of strongly agree to strongly disagree, should someone that has been to half of the countries of the world respond with the midpoint of the scale? Arguably, the answer is likely yes. This represents another feature of these types of items in that they potentially represent a spectrum of behavior with this as the extreme anchor.

We can also examine some of the other items in the study as a potential contrast to these. While we do not have reasonable data to calculate something like a factor analysis on these responses (nor do we think such an analysis would be reasonable in this case), we can think about these items in the light of false positives to group them in ways that they may be similar. Table 4 is the authors' attempt at such a grouping, based on how individuals may have arrived as false positives on these items.

The first two categories are those described above, both representing different types of truths and impossibilities. The first of these two are what we would classify as simple known truths – truths that are simple, clear, and have little to no room for any type of counter-argument. These truths have no gray area, and largely represent a clean no/yes take on the content. The second of these two categories are what we would classify as individual truths or sliding-scale truths. These are items that may actually be true for some individuals in the larger population. A part of this is the spectrum of behavior they potentially represent. It is likely that few individuals have been to every country in the world, but some people have been to most, and more have been to many. These items thus potentially anchor the extreme behaviors at the ends of the scale (strongly agree -> all countries; strongly disagree -> one country), and allow people to move around between these extremes.

Aside from these categories, we suggest four potential, additional, categories. The simplest of these are those instructed response items which simply tell participants which response to select. There are also items that suffer from double-barreling – the response of 'I'm paid biweekly, just not by leprechauns' to the item 'I am paid biweekly by leprechauns' is a good prototypical example of this potential problem. This is also potentially a problem with items where individuals have to nest parts of an item into others – many participants spoke of knowing what jpgs and firewalls were, just not knowing how they were geological terms. Another group of items are those where individuals may somewhat fairly claim ignorance of an uncertain or unknown truth. 'All my friends say I would make a great poodle' is a good example of this, as some individuals who chose the midpoint of the scale pointed out the simple fact that they didn't know what their friends thought, as they'd never asked them about it. We also included here the items which had context-nonsensical words such as 'parabanjology' in them, as some participants justified a neutral response to these items by saying they couldn't reasonably answer the question being asked.

Finally, a group of items seem to be defined by justification through semantic arguments. Put another way, participants justified responses to these items not by arguing with the raw content of the item, but rather with the meaning of some subset of the words. For instance, some individuals who agreed that they could 'teleport across time and space' did so with the justification that walking was a type of teleportation. In addition, some individuals who agreed that they could 'see into the future' did so with the justification that they have a plan for what they're doing tomorrow, or that we could predict things like tomorrow's weather.

While this is almost certainly not the only potential grouping of these items into categories that could be reasonably argued, we believe that it gives researchers additional guidance both in setting up future studies on this topic as well as on how to potentially craft new items of these type in the future. Broadly, this first category of items representing simple known truths appear to have the lowest false positive rates, while each of these other categories have a greater number of potential ways that unexpected responses can be justified. Using these types of items that do not have potential outs are likely to be the best way to limit this type of error in a study.

As with all research, there are limitations drawn from our choice of sample. While the generalizability of these results is thus

Table 4
Potential groupings of current items.

Simple known truths	**Individual/sliding-scale truths**
I work fourteen months in a year	I have been to every country in the world
I have never used a computer	I sleep less than one hour per night
I work twenty-eight hours in a typical work day	I can run 2 miles in 2 min
I am using a computer currently	I have never brushed my teeth
I am participating in an online study currently	I get almost no sleep every night
A dolphin is an animal	I can run a mile in 4 min
Trees are a source of wood	I own a pet stingray
Fish live in water	I frequently fear that cows are going to take over the world.
Humans eat food	I do not understand a word of English
Oranges are fruit	I can eat as much as a horse
	I eat cement occasionally.[1]
	I will be punished for meeting the requirements of my job.[2]
Uncertain/unknown truths	**Double-barreled truths**
All my friends are aliens	I am paid biweekly by leprechauns
All my friends say I would make a great poodle	I am familiar with computer terms such as filtibly and proxypod
I am interested in pursuing a degree in parabanjology.[3]	I am familiar with geological terms such as jpg and firewall
I am fluent in combinatorial English.[4]	I am interested in pursuing a degree in parabanjology.[3]
	I am fluent in combinatorial English.[4]
Semantic argument truths	**Instructed response**
I can teleport across time and space	Answer with 'Agree' for this item
I have the power to cast spells on others	Answer with 'Disagree' for this item
I can see into the future	Answer with 'Strongly disagree' for this item
I am able to read the minds of others	Answer with 'Strongly agree' for this item
I believe I have supernatural powers	Answer with 'Neither agree nor disagree' for this item
I have never spoken to anyone who was listening	
I eat cement occasionally.[1]	
I will be punished for meeting the requirements of my job.[2]	

Note: [1,2,3,4] – These four items could arguably fit well into either of these categories, so we have included them in both for each item.

limited by the fact that our samples are college students from a small liberal arts college in the Midwest, this exercise was again one of proof-of-concept. These studies show that in this sample individuals can be found that treat these items differently than we might expect. There is nothing particularly unique about this sample that should have led to higher or lower rates of this behavior, and it is difficult to imagine that a sample could be found where this general finding of existence of this behavior would not be seen at all. That said, that possibility does always exist.

In all, one of the main takeaways of this study is the simple fact that these types of items should not be used without consideration of potential participant perspectives that are beyond what was originally considered when creating the item. None of these items are likely to be flawless, but understanding these flaws allows researchers to use them in a more sensible and careful way. These items are not without limitations, but as long as those limitations are considered these items are a useful tool in the detection of carelessly invalid responses.

5. Author note

An earlier version of this work was presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA, April 23–25, 2015.

Declaration of Competing Interest

None.

References

Berry, D. T. R., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review, 11*, 585–598.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*, 317–335.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory 2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Curran, P. G. (2016). Methods for the detecting of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced choice self-description checklist. *Personnel Psychology, 15*, 13–24.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.). *Personality Psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96.

Hargittai, E. (2009). An update on survey measures of web-oriented digital literacy. *Social Science Computer Review, 27*, 130–137.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*, 99–114.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2014). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*, 299–311.

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.

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

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

Orpen, C. (1971). The fakability of the Edwards Personal Preference Schedule in personnel selection. *Personnel Psychology, 24*, 1–4.

Simms, L. J., Goldberg, L. R., Roberts, J. E., Watson, D., Welte, J., & Rotterman, J. H. (2011). Computerized adaptive assessment of personality disorder: Introducing the CAT-PD project. *Journal of Personality Assessment, 93*, 380–389.

Sudman, S., Bradburn, N. M., & Schwarz, N. (Eds.). (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA, USA: Jossey-Bass.