Noncompliant responding

Barends, Ard J.; de Vries, Reinout E.

published in
Personality and Individual Differences
2019

DOI (link to publisher)
10.1016/j.paid.2019.02.015

document version
Publisher's PDF, also known as Version of record

Personality and Individual Differences 143 (2019) 84–89

Contents lists available at ScienceDirect

Personality and Individual Differences

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

Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality

Ard J. Barends[a,*], Reinout E. de Vries[a,b]

[a] Vrije Universiteit Amsterdam, Department of Experimental and Applied Psychology, Institute for Brain and Behavior Amsterdam, 1081 HV, Amsterdam, The Netherlands
[b] University of Twente, Department of Educational Science, 7500 AE, Enschede, The Netherlands

ARTICLE INFO

Keywords:
HEXACO
Noncompliant responding
Online research
MTurk

ABSTRACT

Studies on Amazon Mechanical Turk (MTurk) often include check questions in personality inventories to ensure data quality. However, a subset of MTurk workers may give only meaningful responses to these checks while giving noncompliant responses to the other questions. We demonstrate in an analysis of five MTurk datasets using the statistical approach of Lee and Ashton (2018) that this selectively responsive subset can be detected on the HEXACO personality inventory. Our lower bound estimate is that at least 2% in each sample did not get caught with the check questions while giving noncompliant responses on the personality inventory. Overall, researchers who strive to remove noise due to noncompliant responding may benefit from complementing check questions with a statistical approach.

1. Introduction

In recent years, research on Amazon Mechanical Turk (MTurk) has proliferated (e.g., Amir, Rand, & Gal, 2012; Buhrmester, Talifar, & Gosling, 2018). MTurk is a quick, convenient, and relatively inexpensive platform to collect data online from large demographically diverse samples (Chandler, Mueller, & Paolacci, 2014). Various studies have demonstrated that findings from the lab can be replicated on MTurk (e.g., Amir et al., 2012). However, researchers often question the quality of such data (Chandler et al., 2014). One key criticism is that a significant proportion of workers only participate for the financial reward and, consequently, provide noncompliant responses.[1]

Survey research (and especially personality research) on MTurk may be subject to noncompliant responses due to the number of questions that need to be answered (e.g., 200 questions in the HEXACO-PI-R; Lee & Ashton, 2006). In order to screen out noncompliant respondents, several attention check questions are often included in surveys. The most commonly used are infrequency items (e.g., *I eat cement occasionally*; Huang, Bowling, & Liu, 2015) and instructed response items (e.g., *to monitor quality, please respond with 'neutral' for this item*; Meade & Craig, 2012).

The typical finding from MTurk studies is that roughly 10% of the workers do not pass attention check questions (e.g., Zhao, Ferguson, & Smillie, 2017). The removal of such data is warranted as including this data may result in false positives or false negatives (Curran, 2016). Thus, including attention check questions in MTurk research is justifiable to detect and filter out noncompliant responses.

However, these check questions have advantages and drawbacks. Although infrequency items easily blend in the survey, they may create some confusion because some people interpret infrequency questions figuratively (Curran & Hauser, 2018). For instance, these authors found some people agreed with the question *'I get paid by leprechauns bi-weekly'* if they got paid bi-weekly. This confusion requires researcher to be somewhat lenient when flagging responses on infrequency items as noncompliant because they otherwise throw away valuable data (Kim, McCabe, Yamasaki, Louie, & King, 2018).

An advantage of instructed response items is that they are unlikely to cause confusion (Kim et al., 2018). However, Curran and Hauser (2018) found that up to 11% of respondents gave a wrong response to these questions. Consequently, the zero-tolerance approach also seems inappropriate when flagging responses to these questions. Finally, a potential issue for both types of check questions is that including too many of them may annoy respondents, so researchers are advised to limit their use to one check question per 50–100 items (Meade & Craig, 2012; cf. Marjanovic, Bajkov, & MacDonald, 2018).

Some researchers may have refused payment to the participants

* Corresponding author at: Department of Experimental and Applied Psychology, Institute for Brain and Behavior Amsterdam, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
E-mail address: a.j.barends@vu.nl (A.J. Barends).

[1] Other terms are also used to refer to the same phenomenon: careless or inattentive responding (Meade & Craig, 2012), insufficient effort responding (Huang et al., 2015), or (combined as) careless/insufficient effort responding (Curran, 2016).

https://doi.org/10.1016/j.paid.2019.02.015
Received 21 December 2018; Received in revised form 7 February 2019; Accepted 10 February 2019
Available online 21 February 2019
0191-8869/

who failed these check questions (McInnis, Cosley, Nam, & Leshed, 2016). As McInnis et al. showed, MTurk workers have major concerns about getting their work rejected which results in them not receiving payment and lowering their approval rating which allows MTurk workers to the best paying tasks on MTurk. Then again, these rejections may have inadvertently led some workers (who want to abuse the system for their own benefit) to adapt to this researcher strategy. For instance, MTurk workers share information among each other about the inclusion of attention check questions in studies (Chandler et al., 2014). Furthermore, research has demonstrated that MTurk workers have become more attentive to check questions over time and they are more attentive than online student subject pools (Hauser & Schwarz, 2016). Consequently, these check questions may fail to detect respondents who actively search for check questions while giving noncompliant responses to all other items.

However, there is another method to detect noncompliant responses based on the intraindividual consistencies and inconsistencies on the questions within a personality inventory. Individuals who show too much or too little variation (i.e., too low or too high *SD*) in their responses may be suspected of noncompliance. Respondents who only alternate between 'neutral' and 'agree' will show too little variation in their responses. Similarly, individuals who (after recoding their answers to negatively keyed items) 'strongly agree' with half of the items of a factor scale and 'strongly agree' with the other half will show too much variation. Specifically, assuming a balanced number of reverse keyed items, such a person will have a mean score around the midpoint of the scale but a large standard deviation. Lee and Ashton (2018) used this approach for the HEXACO-100 to filter out noncompliant responses and they derived specific cutoff values (see Section 2.2 for exact the cutoff values). Based on this logic, various similar approaches have been independently developed and validated (Dunn, Heggestad, Shanock, & Theilgard, 2018; Marjanovic, Holden, Struthers, Cribbie, & Greenglass, 2015; Weathers & Bardakci, 2015). Our goal is to test the psychometric characteristics and utility of the Lee and Ashton (2018) approach. This approach has the advantage of being more difficult for participants to circumvent than selectively searching for check questions.

The goal of the current research is to demonstrate that there is a subsample of noncompliant respondents who do not get detected by the attention check items and that these people can be detected by the statistical approach. Furthermore, we will also demonstrate that this subsample adds noise to the data as they have meaningless responses on the personality inventory indicated by low reliabilities. An additional goal is to derive preliminary cutoff values for other HEXACO versions (i.e., HEXACO-60 and HEXACO-208).

2. Methods

2.1. Participants

Five samples of American MTurk workers completed one version of the HEXACO personality inventory as part of a variety of different studies. Only MTurk workers who had completed > 5000 other tasks on MTurk (i.e., Human Intelligence Tasks) and had their work approved in > 95% of their tasks (i.e., the approval rating), were eligible to participate. Three samples (N = 781; M_{age} = 36.65; SD_{age} = 11.44; 51.09% men) completed the HEXACO-96 (Lee & Ashton, 2018) and two samples (N = 981; M_{age} = 39.28; SD_{age} = 12.15; 42.92% men) completed the HEXACO-208 (De Vries, Wawoe, & Holtrop, 2015).

2.2. Materials

2.2.1. HEXACO-96

The HEXACO-96 is an adapted version of the HEXACO-100 (Lee & Ashton, 2018) that measures the six personality traits Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). The only difference with the HEXACO-100 is that the four interstitial Altruism items are omitted in the HEXACO-96. The HEXACO-96 measures the six personality dimensions with 16 items for each trait on a five-point Likert-scale (1 = *strongly disagree* to 5 = *strongly agree*).

2.2.2. HEXACO-208

The HEXACO-208 (De Vries et al., 2015) is an adapted version of the 200 item HEXACO-PI-R (Lee & Ashton, 2006). The HEXACO-208 measures each of the six traits with 32 questions and also two interstitial facets Altruism and Proactivity with eight items each. All items from the HEXACO-96 are included in the HEXACO-208 and answers were provided in the same five-point Likert scale format as in the HEXACO-96.

2.2.3. Infrequency items

Four infrequency items from the Jackson Personality Research Form (e.g., *I have never used a telephone*; Fekken, Holden, Jackson, & Guthrie, 1987) were included at random places in both HEXACO questionnaires.[2] Randomization differed between studies. However, within each study the infrequency items were located at the same place in the inventory. Responses to the infrequency scale items were given on the same five-point Likert scale format as the HEXACO inventory. One of these infrequency questions had to be reverse coded. Responses were scored as 1 if (after reverse coding the negatively keyed item) a participant responded with *'disagree'* or *'strongly disagree'*, otherwise the answer was scored as 0. Afterwards, the zero-one scores were summed. We followed the recommendation of Curran (2016) to use 50% (or more) inaccuracy to flag responses as noncompliant (i.e., flagging scores below three). Note that this is a lenient threshold because the probability of passing this threshold with random responses is 52.48%.

2.2.4. Instructed response items

Four instructed response items (e.g., *this is an attention check; please select 'strongly agree'*) were only included in the HEXACO-208. Again, items were included at random places following the same procedure as for the infrequency scale items. The answer was scored as 1 if someone gave the instructed answer and otherwise as 0 and afterwards, these answers were summed. We applied both the zero-tolerance approach advocated by Kim et al. (2018) and the lenient 50% inaccuracy approach suggested by Curran (2016) to these questions. The zero-tolerance approach flagged a respondent as noncompliant if someone had a score below four. The lenient approach flagged a respondent if the score was 0, 1, or 2, whereas scores of 3 or 4 were not flagged. The probability of passing the threshold with random responses was 18.08% for the lenient threshold and 0.002% for zero-tolerance approach. No other thresholds were investigated (e.g., labeling all scores below 2 as noncompliant).

2.2.5. Statistical approach

To detect noncompliant responses, Lee and Ashton (2018) checked answers that were either largely variable or 'flat' on the same domain scale. Response option overuse was determined by taking the standard deviation of an individual's answers to all HEXACO items (before reverse coding negatively keyed items). For instance, someone who answered 3,3,3,3,2 would get a standard deviation of 0.45 and someone who answered 2,3,3,4,2 would get a standard deviation of 0.84. Note that this procedure is similar to the Intra-individual response variability (IRV) index (Dunn et al., 2018). IRV is conceptually related to the long-string analysis (Curran, 2016) but is easier to calculate and can detect a broader range of noncompliance patterns. Responses were considered overused if a standard deviation of 0.70 or lower was found on this

[2] There is one exception, in one sample (n = 515) the HEXACO-208 was the first half of a survey of 449 questions. The check questions were included at random places in the full survey.

Table 1
Number of responses excluded based on each data quality check approach and the Cronbach alphas for each scale of the HEXACO-96. The results are split between flagged and non-flagged responses and these Cronbach alphas were statistically compared.

	n	Hα	Eα	Xα	Aα	Cα	Oα
1. Total sample	781	0.86	0.84	0.90	0.87	0.88	0.86
2. Sample w/o flagged by any approach[a]	96	0.86	0.85	0.91	0.88	0.86	0.86
3 Flagged by any approach	96	0.14	−0.42	0.44	0.45	0.14	0.36
4 $\chi^2(1)$ difference 2 & 3		59.80**	64.52**	60.54**	43.79**	59.80**	43.66**
5. Δ		1.24	1.54	1.25	1.04	1.24	1.04
6. Sample w/o infrequency flagged[a]	72	0.86	0.84	0.91	0.88	0.86	0.86
7. Infrequency flagged	72	−0.28	−0.48	0.09	0.20	0.10	0.18
8. $\chi^2(1)$ difference 6 & 7		51.22**	45.41**	66.98**	48.20**	46.62**	42.65**
9. Δ		1.52	1.52	1.58	1.30	1.27	1.21
10. Sample w/o statistical flagged[a]	46	0.86	0.84	0.90	0.87	0.86	0.86
11. Statistical flagged	46	0.34	−0.52	0.64	0.59	0.18	0.56
12. $\chi^2(1)$ difference 10 & 11		21.41**	28.75**	15.09**	12.30**	27.00**	12.23**
13. Δ		1.06	1.54	0.88	0.79	1.21	0.78
14. Sample w/o unique infrequency flagged[a]	50	0.86	0.84	0.91	0.88	0.87	0.86
15. Unique infrequency flagged	50	−0.19	−0.35	0.12	0.28	0.05	0.11
16. $\chi^2(1)$ difference 14 & 15		35.32**	31.31**	45.12**	30.11**	36.02**	31.82**
17. Δ		1.47	1.46	1.56	1.23	1.36	1.27
18. Sample w/o unique statistical flagged[a]	24	0.85	0.82	0.89	0.86	0.87	0.85
19. Unique statistical flagged	24	0.40	−0.17	0.76	0.73	−0.07	0.62
20. $\chi^2(1)$ difference 18 & 19		8.90*	13.05**	2.99	2.13	17.59**	4.19*
21. Δ		0.95	1.28	0.53	0.45	1.44	0.64

H: Honesty-Humility; E = Emotionality; X = Extraversion; A = Agreeableness; C = Conscientiousness; O = Openness to Experience.
In the analyses all negative values were set to 0. Effect sizes were calculated with negative values.
Interpretation of effect size Δ is equivalent to Cohen's d (Liu & Weng, 2009).
* $p < .05$.
** $p < .001$.
[a] Sample reliabilities are based on the average values of 5000 random draws of this sample of size n.

index. Response incoherence was determined by first reverse coding all negatively keyed items and subsequently calculating the standard deviation for each of the six domain factor scales. These six standard deviations were then averaged, and people who scored 1.60 or greater were considered to have incoherent responses. Note that this approach is similar to the inter-item standard deviation (ISD) procedure of Marjanovic et al. (2015). However, Marjanovic et al. used a much lower cutoff for the ISD (i.e., 1.22) that allowed them to discriminate between random and non-random data. These cutoff values were based on the most extreme responses in a student sample (< 0.5% of the sample; specifically, the IRV cutoff was z = −2.6 and ISD cutoff was z = 3.3) that were assumed not to be indicative of meaningful responses (Ashton, personal communication). Respondents that exceeded one or both of these cutoff values were flagged. In the HEXACO-208 samples, only the noncompliance to the HEXACO-96 items was analyzed because no cutoff values were available for this version (Lee, personal communication).

3. Results

All datasets are available via https://osf.io/vrsb2/?view_only=2d8846cb51974028981d5c1fe4f6d3fb.

3.1. HEXACO-96

In order to determine the prevalence of noncompliant responding, the infrequency approach and the statistical approach were used to flag noncompliant responses. In the samples that completed the HEXACO-96, 96 participants (12.29% of the sample) were flagged by at least one method (see Table 1). Of these flagged responses, 50 (6.40% of the sample) were uniquely flagged by the infrequency scale and 24 (3.07% of the sample) by the statistical approach. The classifications (flagged versus non-flagged) by the approaches were significantly correlated ($r = 0.67$, $p < .001$).

Table 1 shows that the responses flagged by any of the approaches were unusable based on their Cronbach alphas. Responses flagged by the infrequency scale had alphas between −0.48 and 0.20. Similarly, responses flagged by the statistical approach had alphas between −0.52 and 0.64. Table 1 shows that similar results were also found for the responses uniquely flagged by the two approaches.

The differences in the alphas between those that were flagged and not flagged were statistically compared with the cocron web-interface (Diedenhofen & Musch, 2016). The cocron interface uses an R script to test whether two Cronbach's alphas are derived from the same distribution. This distribution approximates a χ^2 distribution under a true null hypothesis. The same number of respondents were included in the comparison as the number of flagged respondents. This comparison sample was randomly drawn from the group that was not flagged by the same approach. This procedure was repeated 5000 times and the average alphas of these samples were used for the comparisons. As Table 1 shows, all except two comparisons were significantly different and the effect sizes were often large.[3] These results indicate that those flagged by any of the approaches (jointly and uniquely) had low quality data. Removing the flagged data increased alphas often by 0.01 or 0.02 in the retained sample (see Table S2). For example, the full sample had an Agreeableness alpha of 0.87. When the data from 96 respondents flagged by any approach were removed from this sample (Agreeableness alpha = 0.45 of this subsample), the retained sample of 685 respondents had an alpha of 0.89 on Agreeableness. Using the Spearman-Brown prophecy formula (Brown, 1910), we found that this increase in reliability is equivalent to adding three additional items to the subscale.

3.2. HEXACO-208

Similarly, in the samples that completed the HEXACO-208, the infrequency scale, the instructed response scale (using the zero-tolerance approach), and the statistical approach were compared. In total 168

[3] The effect sizes were calculated following the formula of Liu and Weng (2009). Note that Liu and Weng found that with small samples these estimates may be slightly less accurate. Therefore, the effect sizes should be interpreted with this caveat in mind.

Table 2
Number of responses excluded based on each data quality check approach and the Cronbach alphas for each scale of the HEXACO-208. The results are split between flagged and non-flagged responses and these Cronbach alphas were statistically compared.

	n	Hα	Eα	Xα	Aα	Cα	Oα
1. Total sample	981	0.94	0.88	0.93	0.91	0.91	0.89
2. Sample w/o flagged by any approach[a]	168	0.93	0.90	0.95	0.93	0.92	0.91
3. Flagged by any approach	168	0.76	0.20	0.41	0.22	0.59	0.59
4. $\chi^2(1)$ difference 2 & 3		55.60**	141.14**	186.01**	179.31**	93.14**	81.45**
5. Δ		0.86	1.45	1.72	1.68	1.14	1.06
6. Sample w/o infrequency flagged[a]	127	0.93	0.90	0.94	0.92	0.91	0.91
7. Infrequency flagged	127	0.70	−0.30	−0.25	−0.25	0.24	0.27
8. $\chi^2(1)$ difference 6 & 7		57.03**	125.76**	170.93**	145.43**	111.10**	107.65**
9. Δ		1.01	1.79	2.11	1.91	1.48	1.46
10. Sample w/o statistical flagged[a]	74	0.93	0.89	0.94	0.91	0.91	0.90
11. Statistical flagged	74	0.78	0.05	0.61	0.18	0.58	0.73
12. $\chi^2(1)$ difference 10 & 11		21.17**	65.46**	51.61**	68.14**	36.61**	16.16**
13. Δ		0.80	1.50	1.30	1.54	1.07	0.69
14. Sample w/o instructed flagged[a]	87	0.93	0.89	0.94	0.91	0.91	0.90
15. Instructed flagged	87	0.73	0.36	0.24	0.18	0.56	0.33
16. $\chi^2(1)$ difference 14 & 15		33.91**	54.70**	100.04**	80.28**	45.51**	62.51**
17. Δ		0.94	1.23	1.77	1.54	1.10	1.32
18. Sample w/o unique infrequency flagged[a]	38	0.93	0.88	0.93	0.90	0.90	0.89
19. Unique infrequency flagged	38	0.77	−0.18	−0.90	0.18	0.54	0.16
20. $\chi^2(1)$ difference 18 & 19		11.52**	32.25**	46.13**	31.85**	18.23**	30.07**
21. Δ		0.83	1.59	2.30	1.46	1.06	1.41
22. Sample w/o unique statistical flagged[a]	18	0.93	0.86	0.92	0.89	0.90	0.88
23. Unique statistical flagged	18	0.79	0.22	0.73	0.56	0.78	0.88
24. $\chi^2(1)$ difference 22 & 23		4.55*	10.32*	5.51*	7.02*	2.40	0.00
25. Δ		0.76	1.20	0.85	0.96	0.55	0.00
26. Sample w/o unique instructed flagged[a]	21	0.93	0.86	0.92	0.89	0.91	0.88
27. Unique instructed flagged	21	0.71	0.77	0.68	0.67	0.80	0.60
28. $\chi^2(1)$ difference 26 & 27		8.65*	1.14	8.26*	5.35*	2.90	6.36*
29. Δ		0.99	0.35	0.96	0.76	0.56	0.84

H: Honesty-Humility; E = Emotionality; X = Extraversion; A = Agreeableness; C = Conscientiousness; O = Openness to Experience.
In the analyses all negative values were set to 0. Effect sizes were calculated with negative values.
Interpretation of effect size Δ is equivalent to Cohen's d (Liu & Weng, 2009).
* $p < .05$.
** $p < .001$.
[a] Sample reliabilities are based on the average values of 5000 random draws of this sample of size n.

participants (17.13% of the sample) were flagged as giving non-compliant responses by at least one method. Of these participants, 91 were identified by two or more approaches and 77 respondents (7.85%) were uniquely identified by only one approach. In absolute numbers, the infrequency approach flagged 127 respondents (12.95%; 38 uniquely [3.87%]), the instructed response questions flagged 87 respondents (8.87%; 21 uniquely [2.14%]), and the statistical approach flagged 74 respondents (7.54%; 18 uniquely [1.83%]). Again, the classifications by the statistical approach and the infrequency scale were significantly correlated ($r = 0.51$, $p < .001$) as were the classifications by the infrequency scale and the instructed response scale ($r = 0.56$, $p < .001$), and classifications by the statistical approach and the instructed response scale ($r = 0.33$, $p < .001$).

Table 2 shows that HEXACO scales in samples consisting of respondents who were flagged by any of the approaches had low reliabilities. Again, these Cronbach alphas of the flagged responses were statistically compared to the rest of the sample using the procedure described above. The results demonstrated that in 38 of the 42 comparisons the alphas of the flagged responses were significantly lower than those of the non-flagged responses.

Furthermore, the effect sizes were often very large (> 1). Again, removing these flagged responses improved the alphas in the retained sample by 0.01 or 0.02 (see Table S3).

If the instructed response questions were flagged at the lenient (50% inaccuracy) threshold advised by Curran (2016), instead of the zero-tolerance threshold, then the overall number of detections dropped to 152 responses (15.49% of the sample). This leniency hampered both the overall and unique ability of the instructed response questions to detect noncompliant responses (32 [3.26%] and five responses [0.51%] respectively).

Comparatively, the infrequency scale now uniquely flagged 61 respondents (6.22%) and the statistical approach uniquely flagged 19 respondents (1.94%). Again, Table S1 shows that the responses flagged by these approaches had significantly lower alphas than the non-flagged responses. Only the five responses uniquely flagged by the instructed response scale were not significantly different from the non-flagged items.

3.3. Preliminary cutoff values

In order to determine preliminary cutoff values for detecting non-compliant responses for the shortest and the longest versions of the HEXACO personality inventory (HEXACO-60 and HEXACO-208 respectively) we applied the same criteria of the HEXACO-100 to the data using the items of these other versions. That is, response overuse and incoherence were determined among the 60 items only using the same cutoff values (i.e., IRV < 0.70 and ISD > 1.60). Similarly, we employed this procedure among all 208 items using again the same cutoff values.

The cutoff values of the HEXACO-100 worked remarkably well when applied to all 208 items: 71 of the 74 (95.95%) flagged responses were again identified. Note that two of the three missed responses were also flagged by the check questions. Therefore, close to 99% of the same participants were again flagged by any of the methods. Furthermore, nine responses were now additionally flagged by the statistical approach (two were also flagged by the check questions).

When only including the items of the HEXACO-60 in the statistical analysis (with the data aggregated across all five datasets) 105 of 120 (87.50%) of the flagged responses were also flagged when using the above cutoff values. Furthermore, 13 of the missed responses were

flagged by the check questions. Therefore, also for the HEXACO-60 close to 99% of the participants were flagged again by any of the methods. Additionally, 10 extra responses were also flagged by the statistical approach (two of these were flagged by the attention check questions).

4. Discussion

Our results show that in MTurk samples roughly 15% of respondents were flagged for noncompliant responses. This portion of noncompliant respondents on the HEXACO-PI-R is greater than found in other types of samples (e.g., 0.5% of students and 1% of the online hexaco.org respondents; Ashton, personal communication). Furthermore, there was a small, but significant, subsample of workers who actively searched for the check questions and only selectively responded in a meaningful way to these questions while they gave noncompliant responses to the other questions. However, this subsample could be detected with the statistical approach of Lee and Ashton (2018). All flagged responses clearly were unusable as reliabilities of these samples were often low or even negative and, in almost all instances, these reliabilities were significantly different from similar sized compliant samples.

Our results also show that the infrequency scale, instructed response questions, and the statistical approach each had their own unique ability to detect noncompliant responses. This is not surprising as correlations between detection methods are far from perfect (i.e., 0.33 to 0.67 in our samples; see also DeSimone & Harms, 2017; Niessen, Meijer, & Tendeiro, 2016). Moreover, the chance to pass the attention check questions with random responses was rather high (around 20% for the instructed response scale and over 50% for the infrequency scale). Therefore, our findings clearly support the recommendation of Curran (2016) to combine multiple approaches to detect noncompliant responses.

The statistical approach can also compensate for legitimate concerns about the use of intuitive—but overly stringent—zero-tolerance exclusion criteria to attention check questions (Curran & Hauser, 2018). Our findings show that when the recommended lenient approach was used to flag responses on the instructed response questions (e.g., Curran, 2016), the number of flagged responses dropped markedly. However, the infrequency scale and the statistical approach partially compensated for this drop as only 16 fewer respondents were flagged overall. Furthermore, note that we included four times as many check questions as advised by Meade and Craig (2012). Consequently, the statistical approach may work even better for researchers who worry about potentially annoying participants with too many check questions (but see Marjanovic et al., 2018).

Although the statistical approach of Lee and Ashton (2018) was developed for the HEXACO-100, we found these cutoff values to also perform adequately among other versions of the HEXACO Personality Inventory (e.g., HEXACO-60; Ashton & Lee, 2009). That is, the very same cutoff values consistently flagged the same respondents in the 60 and 208 item versions. Note, however, that these cutoff values are preliminary and that it is yet unclear whether they can be included in other personality inventories. However, we expect that these criteria are also applicable to other personality inventories with highly similar item content and response formats such as the NEO-PI-R (Costa & McCrae, 1992).

4.1. Limitations and directions for future research

We want to stress that a statistical approach is not a magic bullet to detect all forms of noncompliant responses. We found a less than perfect overlap in detection by the check questions and the statistical approach, suggesting that different approaches are needed to detect all forms of noncompliant responding. Moreover, it seems valuable to further investigate how additional data-screening methods such as the odd-even consistency approach and long-string analysis (Meade & Craig, 2012) complement the statistical, infrequency, and instructed response approaches.

Some other limitations to the current use of statistical approach can be noted. Most importantly, our MTurk samples did not allow us to investigate the occurrence rate of Type I and Type II errors. Our estimated prevalence of workers who actively search for check questions is likely a lower bound estimate. The statistical approach was somewhat lenient as it clearly did not flag all noncompliant responses. Therefore, the statistical approach requires further validation, those who want to use this approach need to proceed with caution. Note that the cutoff values were derived from one specific sample and it may be appropriate to test whether these cutoff values also generalize to other samples. Furthermore, the statistical approach could also be compared to similar approaches and cutoff values (e.g., Dunn et al., 2018; Marjanovic et al., 2015; Weathers & Bardakci, 2015). Therefore, future research may want to compare how different cutoff values are able to correctly classify samples instructed to give noncompliant responses (while trying to avoid detection) from samples instructed to give honest responses.

Our results are also informative for preregistration procedures. As noted in the introduction, depending on the source and the type of check question, lenient or a zero-tolerance thresholds are suggested to flag responses on attention check questions (Curran, 2016; Curran & Hauser, 2018; Kim et al., 2018). This may have led researchers to often apply their own procedures to exclude responses (or similarly, to not exclude any responses). Similarly, it seems that researchers often do not communicate their data cleaning practices of MTurk studies (Chandler et al., 2014). Therefore, in order to further increase transparency and advance best practices, researchers are advised to carefully consider beforehand how they will ensure data quality and preregister such procedures before they start data collection. This seems particularly useful for MTurk research because it has relatively high rates of noncompliance (although these concerns may also apply to other online samples; see Thomas & Clifford, 2017).

A limitation of the current study is that the impact of the exclusions on external validity (e.g., the relation between a personality trait and a dependent variable) could not be tested. The attention check questions and the statistical approach were used to screen out MTurk workers for the follow-up studies that included validity measures. Therefore, it is unclear whether retaining or excluding data affects statistical inferences or substantially influences effect size estimates. Future research should investigate whether these exclusion procedures also improve external validity.

The current findings likely generalize to other personality and survey research on MTurk. It is an outstanding question whether selective responding to check questions is also an issue on other crowdsourcing platforms. We expect that our findings will generalize to platforms where passing check questions is incentivized. For instance, these concerns may also apply to online student participant pools if credit is withheld based on the failure to answer correctly to check questions. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

4.2. Conclusion

In conclusion, our findings show there is a proportion of selectively responsive MTurk workers and thus only relying on attention check questions is not optimal to guarantee data quality. Filtering out these respondents will likely improve data quality. Researchers are advised to carefully consider and combine multiple data quality checks when planning a research project using online crowdsourcing platforms.

Funding

This research was financially supported by a grant from LTP business psychologists to support a PhD-position for the first author at the VU University, The Netherlands.

Acknowledgments

We would like to thank Michael C. Ashton, Kibeom Lee, and Catherine Molho for their feedback on earlier draft of this manuscript. We also want to thank Joost Jongeneel, Lisanne Kahlman, and Sanne Veefkind for their help with the data collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.paid.2019.02.015.

References

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the internet: The effect of $1 stakes. *PLoS ONE, 7*(2), e31461.

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure for the major dimensions of personality. *Journal of Personality Assessment, 91*(4), 340–345.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322.

Buhrmester, M. D., Talifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science, 13*(2), 149–154.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods, 46*(1), 112–130.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

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

Curran, P. G., & Hauser, K. A. (2018). *I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention checks.* (Manuscript under review).

De Vries, R. E., Wawoe, K. W., & Holtrop, D. (2015). What is engagement? Proactivity as the missing link in het HEXACO model of personality. *Journal of Personality, 84*(2), 178–193.

DeSimone, J. A., & Harms, P. D. (2017). Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business Psychology, 33*(5), 559–577.

Diedenhofen, B., & Musch, J. (2016). cocron: A web interfae and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science, 11*(1), 51–60.

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparisons to other indicators and relationships with individual differences. *Journal of Business Psychology, 33*, 105–121.

Fekken, G. C., Holden, R. R., Jackson, D. N., & Guthrie, G. M. (1987). An evaluation of the personality research form with Filipino university students. *International Journal of Psychology, 22*, 399–407.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavioral Research Methods, 48*, 400–407.

Huang, J. L., Bowling, N. A., & Liu, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business Psychology, 30*(2), 299–311.

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

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO personality inventory: Two new facet scales and an observer report form. *Psychological Assessment, 18*(2), 182–191.

Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment, 25*(5), 543–556.

Liu, H.-Y., & Weng, L.-J. (2009). An effect size index for comparing two independent alpha coefficients. *British Journal of Mathematical and Statistical Psychology, 62*, 385–400.

Marjanovic, Z., Bajkov, L., & MacDonald, J. (2018). The conscientious responders scale helps researchers verify the integrity of personality questionnaire data. *Psychological Reports*. https://doi.org/10.1177/0033294118783917.

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

McInnis, B., Cosley, D., Nam, C., & Leshed, G. (2016). Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2271–2282). ACM.

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

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

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior, 77*, 184–197.

Weathers, D., & Bardakci, A. (2015). Can response variance effectively identify careless respondents to multi-item, unidimensional scales? *Journal of Marketing Analytics, 3*(2), 96–107.

Zhao, K., Ferguson, E., & Smillie, L. D. (2017). Individual differences in good manners rather than compassion predict fair allocations of wealth in the dictator game. *Journal of Personality, 85*(2), 244–256.