Using Markov Chains to Detect Careless Responding in Survey Research

Torsten Biemann[1], Irmela Koch-Bayram[1], Madleen Meier-Barthold[2] & Herman Aguinis[3]

[1] University of Mannheim, Germany
[2] Rotterdam School of Management, Erasmus University, the Netherlands
[3] Department of Management, The George Washington University, USA

Author Note

Torsten Biemann https://orcid.org/0000-0003-1728-6765
Irmela Koch-Bayram https://orcid.org/0000-0002-8924-1235
Madleen Meier-Barthold https://orcid.org/0000-0002-6617-3436
Herman Aguinis https://orcid.org/0000-0002-3485-9484

Torsten Biemann is Professor of Human Resource Management and Leadership at the University of Mannheim, Germany. His research focuses on methods in management research, HRM systems, careers and career patterns, and leadership effectiveness.

Irmela F. Koch-Bayram is an Assistant Professor at the Chair of Business Administration, Human Resource Management and Leadership, University of Mannheim, Germany. Her research examines employee perceptions and behavior in the context of corporate social responsibility, as well as human-AI collaboration and decision-making.

Madleen Meier-Barthold is an Assistant Professor of Human Resource Management at Rotterdam School of Management, Erasmus University Rotterdam, in the Netherlands. Her research focuses on the strategic role of HRM in organizations, investigating how HR policies, practices, and leadership shape employee perceptions, decision-making, and organizational outcomes. Her work employs a wide range of advanced quantitative methods and has been published in leading journals such as *Human Resource Management*.

Herman Aguinis is the Avram Tucker Distinguished Scholar and Professor of Management at The George Washington University School of Business. His work focuses on the global acquisition and deployment of talent in organizations and organizational research methods (i.e., behavioral science and data science). Every year since 2018, Web of Science Highly Cited Researchers Reports has ranked him among the world's 100 most impactful researchers in Economics and Business, he served as President of the Academy of Management, and has been inducted into The PhD Project Hall of Fame. He has published 13 books, including *Research Methodology: Best Practices for Rigorous, Credible, and Impactful Research* (2025) and over 220 articles in refereed journals. He is a fellow of the Academy of Management (AOM), former editor of *Organizational Research Methods*, and received the Academy of Management Research Methods Division Distinguished Career Award. For more information, visit www.hermanaguinis.com.

Using Markov Chains to Detect Careless Responding in Survey Research

Abstract

Careless responses by survey participants threaten data quality and lead to misleading substantive conclusions that result in theory and practice derailments. Prior research developed valuable precautionary and post-hoc approaches to detect certain types of careless responding. However, existing approaches fail to detect certain repeated response patterns, such as diagonal-lining and alternating responses. Moreover, some existing approaches risk falsely flagging careful response patterns. To address these challenges, we developed a methodological advancement based on first-order Markov chains called *Lazy Respondents* (Laz.R) that relies on predicting careless responses based on prior responses. We analyzed two large datasets and conducted an experimental study to compare careless responding indices to Laz.R and provide evidence that its use improves validity. To facilitate the use of Laz.R, we describe a procedure for establishing sample-specific cutoff values for careless respondents using the "kneedle algorithm" and make an R Shiny application available to produce all calculations. We expect that using Laz.R in combination with other approaches will help mitigate the threat of careless responses and improve the accuracy of substantive conclusions in future research.

Keywords: Careless responding, data screening, data cleaning, Markov chains, surveys

Using Markov Chains to Detect Careless Responding in Survey Research

Data collection through surveys is subject to considerable data quality threats because of careless response behaviors (DeSimone et al., 2015). Careless responding, also referred to as random response (Beach, 1989), insufficient effort responding (Huang et al., 2012; Huang & DeSimone, 2021), or inattentive responding (Curran, 2016), occurs when participants respond to survey questions without paying sufficient attention to the items or the instructions (Meade & Craig, 2012). Careless responding leads to serious consequences such as distorted means and covariance structures, item correlations, factor loadings, and construct dimensionality (Arias et al., 2020; Goldammer et al., 2020; Kam, 2019; Kam & Meyer, 2015; Huang et al., 2015b). Problems caused by careless responding are not just mere methodological curiosities. Careless respondents damage scales' reliability, leading researchers and practitioners to rely on distorted and imprecise measures (Arias et al., 2020; Huang et al., 2012; Maniaci & Rogge, 2014). In addition, because it inflates and/or deflates Type I error rates, careless responding leads to over- and underestimation of the strength of relations between variables (e.g., Goldammer et al., 2020; Huang et al., 2015b; Maniaci & Rogge, 2014).

Careless responding can be broadly categorized into *random responding* and *non-random patterned responding*, with the latter including straightlining (i.e., identical consecutive responses) and seesaw responding, which comprises repeated response patterns such as diagonal-lining (e.g., 1-2-3-4-5-4-3-2), alternating extreme pole responses (e.g., 5-1-5-1-5-1-5-1) and alternating responses with low variance (e.g., 4-5-4-5-4-5-4-5) (DeSimone et al., 2018; Meade & Craig, 2012; Ulitzsch et al., 2022). Unfortunately, there is abundant evidence that careless responding is pervasive in organizational survey research (Goldammer et al., 2020). For example, base rates of careless responding range from 3.5% (Johnson, 2005) to 10.6% (Kurtz & Parrish, 2001) and even more than 50% (Baer et al., 1997), depending on how careless responding is

measured. Given the increased popularity of online surveys (e.g., Ward & Meade, 2023), addressing careless responding has become an even more urgent methodological challenge.

Prior research has developed several robust precautionary (e.g., page time, instructed response items) and post-hoc procedures to detect careless responding (as described in the next section of our article). While we acknowledge that proposed solutions developed to date are undoubtedly helpful in detecting a specific type of careless responding, we provide evidence that existing approaches fail to detect other careless response patterns or falsely flag careful response patterns that are particularly pernicious in leading to incorrect substantive conclusions. Therefore, we developed a post-hoc procedure to detect patterned careless responding: the *Lazy Respondents* (Laz.R) index. Our Laz.R index differs from existing solutions because we use first-order Markov chains to measure the degree to which participants display *careless response behavior*. This methodological innovation relies on the premise that, for careless respondents, the last response is a useful predictor of the following response. Consequently, our approach is more powerful in detecting non-random patterned responding, such as straightlining and seesaw responding, as it employs a broader definition of patterned responding and serves as a valuable addition to other procedures.

The remainder of our article is structured as follows. First, we briefly overview existing precautionary and post-hoc measures of careless responding and highlight their limitations. Next, we describe the development of Laz.R and compare the reliability and validity provided by careful and careless respondents as identified by Laz.R. We also compare Laz.R to the most widely used precautionary and post-hoc measures of careless responding in large, publicly available datasets as well as in an experimental study. By doing so, we provide empirical evidence that Laz.R improves validity compared to existing methodological approaches. Specifically, we ascertained that Laz.R correctly identifies patterns of careless respondents that

some or all other indices overlook. In addition, we describe how to use the "kneedle algorithm" (Satopää et al., 2011) to identify sample-specific cutoff values needed to distinguish careful versus careless respondents. Lastly, we provide specific recommendations on using Laz.R and introduce a user-friendly R Shiny application to detect careless respondents in future research.

Existing Approaches for Detecting Careless Respondents and their Limitations

Approaches for detecting careless respondents can be classified into two main types (Curran, 2016; Meade & Craig, 2012). First, *precautionary approaches* to prevent careless response behaviors through survey design choices, such as including specific items or scales, which capture a variety of careless response behaviors but are not always feasible or available (e.g., when researchers use archival data; Hill et al., 2022). On the other hand, *post-hoc approaches*, sometimes labeled indirect measures (Goldammer et al., 2020), are based on conducting analyses after data collection and usually analyze response patterns based on item content or order. These approaches use different logical concepts, including pattern indices, outlier analysis, and consistency indices, to detect cases of content non-responsivity.

Precautionary Approaches

This section summarizes the most common precautionary measures. More extensive discussion of these and other approaches can be found in Bowling et al., 2023; Curran, 2016; DeSimone et al., 2015; Meade & Craig, 2012; and Ward & Meade, 2023.[1]

Response Time

This technique excludes respondents based on a minimum response time for the entire survey, survey pages, or single items. The rationale is that a minimum amount of time is needed for "careful" respondents to cognitively process the questionnaire items, recall response-relevant

[1] In addition to the more typical precautionary approaches described in this section, there are some innovative approaches, like eye tracking or visual elements of online questionnaires (Ward & Meade, 2023).

information, and translate this information into a response (Bowling et al., 2023; Huang et al., 2012).

Self-reported Indicators

Self-reported indicators are attention-check questions that ask participants whether they answered carefully and honestly or whether they paid sufficient attention or devoted effort to the study (e.g., "I carefully read every survey item," Meade & Craig, 2012).

Infrequency Items

Infrequency items, also known as bogus items, prompt for unambiguous correct or incorrect responses (e.g., "I have never used a computer," Huang et al., 2015a). These items can be used at several points throughout the survey to identify participants who fail one or more items (Durran, 2016; DeSimone et al., 2015).

Instructed Response Items (IRIs)

IRIs are items such as "respond with strongly disagree for this item." These items have the advantage that answers are clearly instructed and give no leeway for expected responses (Meade & Craig, 2012). Like infrequency items, these questions can be used at several points throughout the survey to screen out participants who failed at least once.

Inability to Recognize Item Content

Bowling et al. (2023) argued that careless respondents are less likely to read and process the content of a questionnaire thoroughly and, thus, would be less likely to recognize item content. The authors asked survey participants to respond to ten multiple-choice questions about the content of items they previously included in their main survey.

While the precautionary approaches are useful and valuable, they also have limitations. For example, participants often perceive self-reported indicators, infrequency items, and instructed response items as insulting, especially when they voluntarily participate in the study.

Crowdsourcing platform participants (e.g., MTurk) are quite familiar with such items and, thus, easily detect and correctly respond to them even though they do not carefully respond to each item throughout the questionnaire. Moreover, researchers often find themselves in situations that do not allow for changes in survey design ex-post (e.g., when they use secondary data sources) or require further statistical steps to detect careless respondents after data collection. Therefore, post-hoc measures, which we describe next, are essential.

Post-hoc Approaches

Longstring Index

This technique builds on the idea that careless respondents resort to answering patterns in which they repeatedly choose an identical answer option, called strings (Johnson, 2005; Costa & McCrae, 2008). Based on the computation of either the *maximum string* or the *average string* length, cases of careless responses are identified (Meade & Craig, 2012). For example, an answer pattern of 1-1-1-1-1-2-2 on a scale from 1 = *disagree* to 5 = *agree* includes a string of length five (1-1-1-1-1) and a string of length two (2-2), resulting in a maximum string of 5 and an average string of 3.5. The longstring index is limited in terms of the types of detectable patterns. For example, it can detect straightlining but not cases where respondents chose any other pattern, like seesaw response patterns (e.g., 1-2-3-4-5-4-3-2-1 or 1-2-1-2-1-2-1).

Mahalanobis Distance

As an outlier index (Aguinis et al., 2013), the Mahalanobis distance (D; Mahalanobis, 1936) has also been suggested to detect careless respondents (Meade & Craig, 2012). Mahalanobis D flags respondents who respond substantially atypically compared to others in the sample. It is computed as the multivariate distance between a respondent's response vector and the vector of the sample means. High values indicate high deviances from the sample mean on

Mahalanobis D, and further attention is needed since they might be careless respondents. There is no clear cutoff, but outliers can be detected with the help of a quantile plot (as implemented by the R *careless* package). Clearly, careless responding is not the only reason why a respondent might deviate from typical respondents. Thus, focusing on Mahalanobis D might lead researchers to discard accurately responding and "interesting" outliers (Aguinis et al., 2013).

Intra-individual Response Variability (IRV)

IRV measures an individual's standard deviation of responses across a set of consecutive items (Dunn et al., 2018). The implementation of this index varies across studies. Some propose that careless respondents have a particularly low IRV and thus have a relatively invariant response pattern (e.g., 1-2-1-2-1-2-1; Dunn et al., 2018; Goldammer et al., 2020). If researchers follow the recommendation to flag participants with a low IRV, they cannot detect cases where respondents chose seesaw response patterns (e.g., 1-5-1-5-1-5-1-5 or 1-2-3-4-5-4-3-2-1). In a test of popular indices' effectiveness in detecting careless responses, Goldammer et al. (2020) found that IRV does not perform better than chance.

Psychometric/Semantic Synonyms (PsychSyn) and Antonyms

These consistency indices determine whether participants contradict themselves across item pairs. Specifically, this involves computing within-person correlation across item pairs with a strong positive (negative) sample correlation (threshold $|r| > .6$; Goldberg & Kilkowski, 1985; Johnson, 2005; Meader & Craig, 2012). Responses are considered careless if the correlations are not consistent with the underlying notion of synonymy. These consistency indices are not able to detect cases where careless respondents choose central tendency patterns (e.g., 3-3-3-3-3-3-3). Also, these options require the survey to query related items regarding synonymy. In cases where only a few items that are not highly correlated are included in the survey, consistency indices underperform.

Person-fit statistics for Polytomous Items Using Item Response Theory

Person-fit statistics can be used to identify inconsistent item score patterns within the sample or based on the fit of an item-response theory (IRT) model (Niessen et al., 2016). Overall, IRT models describe the probability of a respondent choosing a specific answer to an item based on their latent traits (e.g., personality or intelligence). These models calculate the likelihood of a respondent with a particular trait level selecting a particular answer option, and a low likelihood indicates an inconsistent item score pattern. Two examples are the nonparametric number of Guttman errors G_{Poly} (Meijer et al., 1994) and the lz_{poly} statistic for polytomous items (Drasgow et al., 1985). Guttman errors occur when a respondent's answers deviate from the expected hierarchical pattern of response options on a polytomous item. The expected hierarchical pattern is based on the popularity of each item step in the sample (e.g., from strongly disagree to disagree). lz_{poly} is defined as the standardized log-likelihood of an item score vector under an IRT model (see Drasgow et al., 1985, for more information on the computation).

Laz.R: Theoretical Background and Computation

Careless respondents choose low-effort routes to complete a survey as quickly as possible. In other words, they click through questions without paying much attention to the item's content. From a theory standpoint, Laz.R identifies low-effort routes through a survey and detects careless response patterns. From a statistical standpoint, Laz.R is based on first-order Markov chains. The usefulness of Markov chains in identifying careless respondents was first pointed out by Stark et al. (2017), but their research focused on dichotomous items with the same answering probabilities.

To illustrate the theory underlying Laz.R, consider a situation involving the development of a personality assessment instrument. Our instrument includes 50 items, each rated on a Likert scale, to keep our illustration simple, with anchors ranging from 1 = *disagree* to 4 = *agree*. Lucy,

our first participant, is the prime example of a straightliner and checks the same scale anchor on the left (*1 = disagree*) for all 50 questions. So, in the data file, Lucy's resulting sequence is "1-1-1-1-1-1-1-1…1". John is also a careless respondent but is slightly more creative and uses a seesaw pattern. He starts with the item anchor on the left, moves stepwise to the right, back to the left, and so forth. Thus, John's sequence of answers is "1-2-3-4-3-2-1-2…1-2". As a third illustrative participant, Emma reads all questions carefully and tries to answer truthfully. She selects scale anchors that reflect her personality best and selects the answers "4-2-2-3-3-4-2-1" for the first eight questions. Her answering behavior does not follow a simple pattern based on item order.

Laz.R uses participants' response patterns to detect careless respondents. It is based on the notion that careless respondents' answers to the next question are contingent on the previous answer. Accordingly, analyzing patterns and transitions across answers reveals whether item sequence or item content determines a participant's response behavior. This consists of three steps (please see the online supplement, "Supporting Materials A: Details on Laz.R Computation" for additional technical information).

Step 1: Transition Matrix

In the first step, we summarize patterns of answers in a *transition matrix* **T**, which indicates the number for each possible transition among the s scale anchors:

$$T = \begin{bmatrix} n_{11} & \cdots & n_{1s} \\ \vdots & \ddots & \vdots \\ n_{s1} & \cdots & n_{ss} \end{bmatrix}. \quad (1)$$

In our example, the element n_{12} indicates the absolute number of cases when a 1 is followed by a 2 in an individual's answer sequence. For example, we find this transition at the beginning of John's answering sequence and whenever he starts counting upwards again.

Step 2: Transition Probability Matrix

Based on the transition matrix, we compute a *transition probability matrix* **P**, which gives the probabilities that a specific answering option is followed by each other answering option. As an example, the transition matrix and transition probability matrix for John is as follows (including all 50 questionnaire items):

$$T_{John} = \begin{bmatrix} 0 & 9 & 0 & 0 \\ 8 & 0 & 8 & 0 \\ 0 & 8 & 0 & 8 \\ 0 & 0 & 8 & 0 \end{bmatrix} \text{ and } P_{John} = \begin{bmatrix} 0 & 1.0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 1.0 & 0 \end{bmatrix} \quad (2)$$

For John, Equation 2 shows that a 1 is always followed by a 2, and a 2 is either followed by a 1 or a 3, which results from his pattern consisting of 1-2-3-4-3-2-1. Accordingly, the transition probabilities for 2 are $p_{21} = 0.5$ and $p_{23} = 0.5$, and $p_{22} = p_{24} = 0$, as the two latter transitions do not occur in John's answering pattern. Mathematically, this can be understood as the first-order Markov chain transition probability because we compute the probability for the following state solely based on the current state. For John's case, it might seem plausible to include not only his latest answer to predict the f answer. If we know the answer that precedes the 2, we know whether he is currently counting upwards or downwards and, hence, if a 1 or a 3 will follow. Note, however, that transition matrices and transition probability matrices for higher-order Markov processes are much more complex, and we will later show that first-order Markov chains can also identify such cases of careless responding.

Step 3: Laz.R Scores

In the next step, we multiply each element of **P** by the respective element in **T** (i.e., Hadamard product $\boldsymbol{P} \circ \boldsymbol{T}$) to measure the predictability of sequence elements. Computing the sum of the matrix elements from all transitions and dividing it by the total number of transitions results in Laz.R, which captures the degree to which the previous answer can predict a respondent's answer. Note that final scores can theoretically range between 1 and 0.25 (for four

answering options as in our illustration—the lowest values vary depending on the number of anchors used) and that higher scores indicate high prediction accuracy of responses, which indicates careless responding. Accordingly, Laz.R scores aim to identify respondents who answered with minimum effort. Specifically,

$$Laz.R = \frac{\sum_{i=1}^{s}\sum_{j=1}^{s}(P \circ T)}{N-1} = \frac{\sum_{i=1}^{s}\sum_{j=1}^{s}\left(\begin{bmatrix} p_{11}*n_{11} & \cdots & p_{1s}*n_{1s} \\ \vdots & \ddots & \vdots \\ p_{s1}*n_{s1} & \cdots & p_{ss}*n_{ss} \end{bmatrix}\right)}{N-1}, \quad (3)$$

where s is the number of scale anchors and N is the number of items.[2] Thus, the Laz.R score for John is the following:

$$Laz.R_{John} = \frac{\sum_{i=1}^{s}\sum_{j=1}^{s}\left(\begin{bmatrix} 0 & 1.0*9 & 0 & 0 \\ 0.5*8 & 0 & 0.5*8 & 0 \\ 0 & 0.5*8 & 0 & 0.5*8 \\ 0 & 0 & 1.0*8 & 0 \end{bmatrix}\right)}{49} = 0.67 \quad (4)$$

John's prediction accuracy is relatively high with $Laz.R_{John} = 0.67$ because of his patterned response behavior. Laz.R scores of careful respondents should be much lower, as their responses are less likely to exhibit discernible patterns (see Supporting Materials A for an extended example).

The example above only addresses a specific case to describe the theory behind the development of Laz.R. So, we provide an extended set of patterned responses in Table 1. For example, seesaw responding with ten items alternating between five scale anchors (i.e., 1234543212) generates a Laz.R score of .67. If this pattern continues over a range of fifty items, the corresponding Laz.R score is .63.

To make a value-added contribution to the literature, it is essential to demonstrate that patterned responses impact the reliability and/or validity of research findings and that our

[2] The following function expresses Equation 3 in R: `Laz.R <- function(x){tr <- table(x[-length(x)], x[-1], useNA = "ifany"); sum(tr^2 / rowSums(tr)) / (length(x)-1)}` and, accordingly, `Laz.R(c(1,2,3,4,5,4,3,2,1,2)) = .67`

proposed index effectively identifies instances of careless responding that are not detected by existing indices. Accordingly, we next describe two studies (i.e., Studies 1-2) examining the reliability and validity of careful and careless respondents identified by Laz.R, followed by two additional studies (i.e., Studies 3-4) comparing Laz.R with existing approaches for detecting careless respondents.

Identification of Careful and Careless Respondents Using Laz.R: Implications for Reliability and Validity

We used two publicly available datasets from https://openpsychometrics.org/_rawdata (for a more detailed description of all the datasets we used, please see the section "Supporting Materials B: Description of Datasets" in the online supplement). Both datasets use common scales in a very large sample, covering a great variety of careless response patterns. We computed Laz.R scores for all participants. Then, we compared those individuals with a very high Laz.R score ("careless respondents") to individuals with medium to low scores ("careful respondents"). Specifically, if validity estimates are worse in the group of careless respondents, this would provide evidence of Laz.R's ability to identify careless response patterns. When assessing reliability estimates, we sought to compare careless and careful respondents, as reliability is likely to increase with straightlining but decrease with seesaw responding and random answering patterns. We conducted all analyses using R, version 4.4.1 (R core team, 2024; code available upon request).

Study 1: Reliability and Validity Using Laz.R with Big 5 Personality Dimensions

Sample and Measures

This dataset includes 1,015,342 observations of the IPIP Big-Five Factor Markers, an inventory of 50 items to assess the Big 5 personality dimensions (Goldberg, 1992). The Big 5 personality dimensions have long been the most prominent way to operationalize personality

characteristics and have been shown to predict employees' attitudes, behavior, performance, and career success (Judge et al., 1999; Van Iddekinge et al., 2009). Each dimension was measured with ten items on a 5-point Likert scale ranging from 1 = *disagree* to 5 = *agree* (including 24 reverse-coded items). The item order in the questionnaire followed the same pattern for all participants, with always one item from each Big 5 dimension in the same order. At the end of the survey, participants were asked to indicate whether they had answered the questions accurately and whether their answers could be stored and used for research. Therefore, data for only these participants are available. In addition, we removed all respondents with incomplete questionnaires, which resulted in our study's N = 874,434.

Results

An exploratory view of the data reveals that the extreme answering patterns discussed in Table 1 rarely occurred in the sample. For example, only 0.039% of respondents (341 out of 874,434) answered all questions with scale anchor 1, and 0.0075% (66 respondents) answered the whole questionnaire consistently with the pattern 1-2-3-4-5-4-3-…. However, the data show that most careless respondents did not strictly follow the same extreme pattern throughout the *whole* questionnaire. Instead, many respondents varied their patterned answering behavior. For example, some respondents started with the same scale anchor for the first items, then moved to count up and down; others started with no clear answering pattern for the first items but alternated between 1s and 5s in later parts of the questionnaire. In the following analyses, we thus study not only cases with the most extreme Laz.R scores but select a proportion of careless respondents that also include individuals with partly patterned answering.

Proportion of Careless Respondents. The proportion of careless respondents may vary across studies. For the sake of simplicity, we set the cutoff value to 5%, which is within the range of rates of careless responding reported in previous research (Baer et al., 1997; Johnson, 2005;

Kurtz & Parish, 2001). Thus, we flagged participants with the 5% highest Laz.R scores as careless respondents (N = 43,722), and the remaining 95% are included in the group of careful respondents (N = 830,712).

Reliability. We computed Cronbach's α and zero-order correlations between all five personality constructs for the careful and careless respondents group. We also compared these results with those reported by Burns et al. (2017) and Ehrhart et al. (2008) because these studies use the same scales to provide reliability estimates and correlations among Big 5 dimensions. Results summarized in Table 2 show that reliability estimates are comparable across groups, varying between α = 0.80 and 0.90 for careful respondents and between α = 0.81 and 0.89 for careless respondents.

Validity. Big 5 personality questionnaires intend to capture independent dimensions of an individual's personality (Barrick & Mount, 1991), and thus, we should expect low correlations among the five subscales, indicating discriminant validity. As seen in Table 2, all correlations among subscales were lower in the group of careful respondents compared to careless respondents. Differences ranged from $\Delta r = .14$ (e.g., Extraversion and Agreeableness) to $\Delta r = .37$ (Conscientiousness and Openness). We used the cocor.indep.groups function from the R package cocor (Diedenhofen & Musch, 2015) to perform significance tests for differences between correlation coefficients in two independent groups; all differences between the groups of careful and careless respondents were significant at $p < .001$. Overall, the mean correlation coefficient among personality dimensions was $r = .13$ in the group of careful respondents and $r = .37$ in the group of careless respondents.

To assess convergent validity, we compared results from careful and careless respondents to findings from two published validation studies that used the same 50 IPIP questionnaire items (Burns et al., 2017; Ehrhart et al., 2008). A comparison of the ten correlations among Big 5

personality dimensions in the two validation studies with the careful/careless groups revealed a mean deviation of $\overline{\Delta r}$ = .09 for the group of careful respondents and $\overline{\Delta r}$ = .18 for the group of careless respondents. Thus, results from the group of careful respondents are closer to the correlation pattern among subdimensions found in previously published studies. In summary, results revealed little differences regarding reliability but substantially better discriminant and convergent validity in the group of careful respondents.

Discussion

The differences in correlations computed from the careless compared to careful respondents subgroup are very large compared to typical correlations reported in organizational research (Bosco et al., 2015). Specifically, Bosco et al. (2015) reported that medium effect sizes range from $|r|$ = .09 to .26. These findings contextualize the differences that we found as large, given that they ranged between Δr = .14 and Δr = .37.

Our analyses identified the group of careless respondents by their high Laz.R scores. We argued that high Laz.R scores indicate minimum effort from respondents, resulting in patterned answering behaviors. One might argue that the opposite (i.e., completely random response behavior) can also indicate careless responding. Beach (1989), for example, used the term random responder to describe what we defined as careless responding. Random responding produces an answer pattern with very low predictability and, hence, very *low* Laz.R scores (see the last rows in Table 1 for examples). Accordingly, the respondents with very low Laz.R scores might also be of interest. If truly random responses characterize this group, scale reliability will be low because inter-item correlations that result from random responses are expected to be zero. Additional group analyses with the 5% *lowest* Laz.R scores revealed that scale reliability ranged between .79 (Openness) and .88 (Extraversion). Furthermore, in eight out of ten cases, correlations among subscales were lower in the group with the 5% *lowest* Laz.R scores compared

to the other 95% of participants. Because the Big 5 personality dimensions should be independent, lower subdimension correlations indicate higher discriminant validity. Overall, we did not find evidence for a lower reliability or discriminant validity in the participants with very low Laz.R scores. Thus, we did not separate the individuals with low Laz.R scores for the remaining analyses.

Study 2: Reliability and Validity Using Laz.R with Holland Occupational Themes (i.e., RIASEC)

Sample and Measures

We used 145,828 survey responses to 48 items based on the Holland Occupational Themes (Holland, 1997), a popular taxonomy of individuals' traits as part of the most established theory of careers and vocational choice. The instrument covers vocational interests that have been shown to predict job performance, turnover, and career choices (Song et al., 2024; Van Iddekinge et al., 2011). The questionnaire comprised six subdimensions: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C), or RIASEC. Each is measured using eight items that describe various tasks. Participants used a 5-point scale ranging from 1 = *dislike* to 5 = *enjoy* to rate how much they would enjoy performing each task (e.g., "Design artwork for magazines" from the Artistic scale and "Give career guidance" from the Social scale). Items were from an item pool developed by Liao et al. (2008). After removing incomplete questionnaires, Study 2's N = 135,764.

Results

Proportion of Careless Respondents. Again, for now, we used a cutoff of 5% to distinguish careful from careless respondents, generating a subgroup of N = 128,976 careful and a subgroup of N = 6,788 careless respondents.

Reliability. Results in Table 3 show that reliability was slightly higher in the group of

careless respondents (Cronbach's α ranging from 0.92 to 0.96) than in the group of careful respondents (Cronbach's α ranging from 0.82 to 0.89). An explanation for this result is the lack of reverse-coded items in the RIASEC questionnaire. In other words, participants with suspicious answering patterns such as 1-1-1… or 5-5-5… generate highly consistent results on all subscales.

Validity. Table 3 also shows results for discriminant validity. The pairs of subscales R-S, I-E, and S-C are especially important because they constitute opposites of Holland Occupational Themes' hexagon (Holland, 1997) and should, therefore, yield low or even negative coefficients. Correlations in the group of careful respondents for these three pairs of subscales were $r_{(R\text{-}S)}$ = .04; $r_{(I\text{-}E)}$ = -.01; and $r_{(S\text{-}C)}$ = -.04. For careless respondents, these correlations were $r_{(R\text{-}S)}$ = .47; $r_{(I\text{-}E)}$ = .51; and $r_{(S\text{-}C)}$ = .60. Again, we conducted significance tests for differences between correlation coefficients in two independent groups using the R package cocor. Differences between correlation coefficients of RIASEC dimensions were significant at $p < .001$ when comparing the careful and careless respondent groups.

To assess convergent validity, we compared the RIASEC correlation patterns of careful and careless respondents to findings from a recently published meta-analysis (Hurtado Rúa et al., 2019). Pairing the six RIASEC subdimensions resulted in 15 unique correlation coefficients between subdimensions, which deviated from meta-analytical findings by $\overline{\Delta r} = .11$ for careful respondents and $\overline{\Delta r} = .36$ for careless respondents (see Table 3). Overall, as a constructive replication of results from Study 1 using the Big 5 dataset, discriminant and convergent validity were higher in the group of careful respondents.

As additional evidence, previous research also analyzed gender differences for the RIASEC interest types, summarized in a meta-analysis by Su et al. (2009). In all six RIASEC subdimensions, the deviation of effect sizes from meta-analytic findings was smaller for careful

respondents than for careless respondents, with a mean deviation of $\overline{\Delta d} = .21$ for careful respondents and $\overline{\Delta d} = .34$ for careless respondents.

Discussion

As a consequence of using Laz.R scores to distinguish between careful and careless respondents, we produced noticeable improvements in psychometric properties and substantive results. Regarding reliability, the non-existence of reverse-coded items seemed to increase Cronbach's alpha in the group of careless respondents. Regarding discriminant validity, correlations among subdimensions were much higher in the group of careless respondents, even when no positive correlation was suggested by theory. Lastly, convergent validity was higher in the group of careful respondents, as the correlation patterns among subdimensions in the group of careful respondents were more similar to results from other studies than the correlation patterns of careless respondents.

Value-added Contributions of Laz.R Above and Beyond Existing Approaches for Detecting Careless Respondents

We describe the results of studies examining the benefits of using Laz.R compared to post-hoc (Study 3) and precautionary (Study 4) approaches. We begin with post-hoc approaches because they are more directly comparable to Laz.R.

Study 3: Empirical Comparison of Laz.R with Post-hoc Approaches

Sample and Measures

For an empirical comparison of the different post-hoc measures, we used the same Big 5 data as in Study 1. In addition to Laz.R, we computed the longstring index, Mahalanobis D, IRV, PsychSyn/PsychAnt, G_{Poly}, and lz_{Poly}. For PsychSyn, we set three thresholds with $r_{(item\ i,\ item\ j)} > .40/.50/.60$. For example, there were 56 correlations with $r > .40$ in the data, which were used to

compute PsychSyn. Similarly, thresholds for PsychAnt were set to $r_{(item\ i,\ item\ j)} < -.40/-.50/-.60$. Results for psychometric antonyms with $r_{(item\ i,\ item\ j)} < -.60$ are not shown because correlations among pairs of items did not pass this threshold in this dataset. We used the careless package (Yentes & Wilhelm, 2021) and the PerFit package (Tendeiro et al., 2016) in R to compute established post-hoc measures.

Results and Discussion

We report results in two sections based on the two types of careless responses discussed in the introduction: (a) non-random patterned responding and (b) random responding. First, we defined answering sequences that we consider typical for straightlining and seesaw responding and computed post-hoc indices for these patterns. Second, we produced random answering sequences and computed the different indices for these random response patterns.

Analysis of Typical Answering Patterns for Careless Respondents. The upper section of Table 4 shows selected straightlining and seesaw answering patterns and corresponding values for the five post-hoc careless responding indices. A direct comparison of index values is not feasible because they have different upper and lower limits. Hence, the table shows percentile ranks for index comparison. Thus, a percentile rank of 1 indicates that the case was among the 1% of answering patterns with the most extreme index scores, and hence, we might flag the respective sequence as stemming from a careless respondent. For Laz.R, longstring, and Mahalanobis D, *high* index values flag careless respondents and, thus, receive low percentile ranks. For IRV and PsychSyn, *low* index values indicate careless respondents; hence, we transformed low index values to low percentile ranks for these indices. Note that for IRV, some researchers have suggested to examine cases with *high* index values (e.g., Marjanovic et al., 2015). Following this approach, cases in our analyses with high percentile rank for IRV should be flagged.

The first typical sequence of careless respondents is, again, the answering sequence 1-1-1-1-1-…. This straigthlining pattern is flagged as an extreme case by Laz.R, longstring, Mahalanobis D, IRV, G_{Poly}, and lz_{Poly} (in the 1st or 2nd percentile). PsychSyn did not flag this pattern (99th percentile) because the responses do not vary; thus, a correlation among those highly correlated items in the full sample could not be computed for this case. An analysis of other straightlining and seesaw patterns in Table 4 reveals that longstring and IRV did not flag seesaw patterns (e.g., 5-1-5-1-5-… and 1-2-3-4-5-4-3-…). Also, Mahalanobis D, G_{Poly}, and lz_{Poly} did not flag patterns with values often close to the scale mean (e.g., 4-4-4-4-4-…). Finally, PsychSyn only flagged the pattern 5-1-5-1-5-… as an extreme case (1st percentile) but did not identify the other patterns of careless respondents.

An analysis of these typical patterns of careless respondents revealed that only Laz.R flagged all pre-defined patterns correctly. At the same time, the other indices showed various weaknesses that logically follow from the respective index construction.

Analysis of Random Responding. In addition to patterned responses, random responding is a second category of careless responding that needs to be identified by careless responding indices (De Simone et al., 2018; Meade & Craig, 2012). When respondents choose answers randomly, their answering sequence contains values that depend neither on item content nor order. Note that there are no typical random response patterns (such as straightlining or seesaw answering), but random answering behavior can easily be simulated. Hence, we conducted additional analyses that extended the Big 5 dataset by N = 1,000 simulated random cases. For example, we simulated cases when each answering option has the same probability (uniform distribution). We computed careless responding indices for the extended datasets with N = 874,434 + 1,000 = 875,434 cases and calculated the mean percentile rank of the 1,000 cases of random responding. A very high or low mean percentile rank for an index indicates its ability to

identify random responding. As seen in the lower part of Table 4, Mahalanobis D, PsychSyn, G_{Poly}, and lz_{Poly} ($r > .40/.50$) showed the best performance. For example, the mean percentile rank for PsychSyn with $r > .40$ was 3.3 when we simulated random responses with a central tendency. Laz.R was not capable of identifying random responding.

In sum, the results provided evidence that Laz.R outperformed other post-hoc indices in identifying all types of patterned responding, but it was less useful in detecting random responding.

Study 4: Empirical Comparison of Laz.R with Precautionary Approaches (CR study)

Precautionary procedures are more direct approaches to identifying careless respondents than post-hoc statistical procedures. Therefore, they are valuable indicators of high index quality when both a precautionary measure and a corresponding post-hoc statistical procedure identify the same individual. Since precautionary measures, by definition, must be implemented before data collection, we conducted a careless respondents study (CR study) specifically designed to compare Laz.R scores with various precautionary approaches and other existing post-hoc measures.

Sample and Measures

To examine careless responding and increase the number of careless respondents in our dataset, we chose a study design aimed to elicit a high proportion of careless responding (e.g., a long and exhausting questionnaire with a tedious design), which is often called an "extreme-groups design" approach (Cortina & DeShon, 1998). We also manipulated survey conditions: 50% of participants received standard instructions (control condition), and 50% were instructed to respond without effort (low effort condition, "Please respond to all questions without effort. In fact, we request that you do so. There is no risk of penalty," following Huang et al., 2012).

We recruited 465 participants via Amazon Mechanical Turk (MTurk). Following best-

practice recommendations (Aguinis et al., 2021; Feitosa et al., 2015), we restricted participation to US individuals with a HIT approval rate greater than 95%. Additionally, we recruited only those currently employed or self-employed individuals, as the questionnaire included several work-related questions. In our sample, 38.06% were female, 61.94% were male, and the average age was 31.37 years (SD = 4.94). The highest education level was distributed as follows: 10.32% high school degree (or similar), 0.86% professional degree, 64.09% Bachelor's degree, 23.87% Master's degree, and 0.86% Doctorate degree. On average, participants completed the questionnaire in 9 minutes and received 0.50 USD for their participation.

As in Study 1, participants first completed 50 items from the IPIP to assess the Big 5 personality dimensions. We then incorporated additional scales commonly used in organizational studies to demonstrate the broad applicability of our findings. Specifically, we assessed participants' satisfaction with work itself and with pay (five items each, Bowling et al., 2018), career satisfaction (five items, Greenhaus et al., 1990), job insecurity (three items, Hellgren et al., 1999), three dimensions of psychological empowerment, namely, meaning, competence, and self-determination (three items each, Spreitzer, 1995), perceived needs-supplies fit and perceived demands-ability job fit (three items each, Cable & DeRue, 2002) as well as subjective/occupational stress (four items, Motowidlo et al., 1986).

In addition, we included several indices to detect careless responding. Within the scales mentioned above, we embedded three infrequency items from Huang et al. (2015a) and three instructed response items (e.g., *respond with "disagree" for this item*). Throughout the survey, we included ten items and asked participants to recognize item content at the end of the questionnaire with ten multiple-choice questions, following the approach of Bowling et al. (2023). In total, participants answered 103 items (including 16 embedded careless responding items) that we used for later analyses of response patterns (for additional details, please see the

section “Supporting Materials C: Scales, Survey Design, and Additional Results for Study 4 (CR Study)” in the online supplement). Unless otherwise specified, scales were measured using a 5-point Likert scale (1 = *disagree*, 5 = *agree*).

We also used the 9-item diligence scale by Meade and Craig (2012) at the end of the questionnaire as a robust assessment of self-reported careless responding. We assessed the page time index and computed several post-hoc indices such as longstring, Mahalanobis D, IRV, psychometric synonyms, and the person-fit statistics G_{Poly} and lz_{poly}.[3]

Results and Discussion

Descriptive statistics and correlations between different measures to identify careless respondents are shown in Table 5. An important observation is the generally low to moderate strength of relationships among most careless responding indices. For example, correlations of the content recognition index with other careless respondent indices range between $|r| = .57$ (for psychological synonyms) and $|r| = .02$ (for G_{Poly}). Similarly, correlations of Laz.R scores with other indices vary between $|r| = .66$ (for IRV) and $|r| = .10$ (for psychological synonyms). These results replicate what other researchers have reported before, i.e., most careless responding indices are not highly correlated (e.g., Meade & Craig, 2012; Goldammer et al., 2020; Ulitzsch et al., 2022). It suggests that researchers might not want to rely on a single index when identifying careless respondents in their data but instead employ two or more potent indices.

Several additional findings are noteworthy. First, we included three instructed response items throughout the questionnaire such as “Please select ‘agree’ for this item.” A closer examination of response patterns suggests that these items were largely ineffective. For instance,

[3] Both person-fit statistics were computed using the PerFit package in R (Tendeiro et al., 2016). We also calculated the normed version of the G_{Poly} (Emons, 2008) as well as $U3_{poly}$ (van der Flier, 1982). Since the correlations among these indices were very high, we decided to include only the results for G_{Poly} and lz_{Poly} to be concise.

several respondents consistently chose the same anchor or a seesaw pattern for all items *except* for the three instructed response items (e.g., "444…4445444…4441444…4443444…"). Interestingly, these participants overlooked the three infrequency items (e.g., "I have never used a computer"), as they continued their patterned responding throughout the infrequency items. It is possible that the incentive of payment for questionnaire completion and MTurkers' familiarity with instructed response items (Aguinis et al., 2021) directed some participants' attention to identify instructed responses, but otherwise ignore item content.

Second, inspecting response patterns indicates that careless respondents tend to select scale anchors on the right side of the scale (i.e., to agree to questions in our questionnaire). Participants exhibiting highly patterned responses almost exclusively selected high-scale anchors, resulting in patterns such as "55555…." or "454545…". Consequently, in an explorative analysis, we found a correlation of $r = .46$ between Laz.R scores and the total sum of all items. The tendency of careless respondents to favor specific scale anchors has been previously documented (Costa & McCrae, 2008). This pattern was also evident in the Big Five dataset used in Study 1, where the number of individuals who consistently selected the same scale anchor for all items was unequally distributed on the scale ranging from 1 ("disagree") to 5 ("agree") with 341, 59, 683, 45, and 544 respondents, respectively.

A third noteworthy finding relates to the low-effort condition in our study. We advised 219 participants to respond to all questions with minimum effort (following Huang et al., 2012). However, this treatment was ineffective[4] and only slightly influenced their response behavior. The last row in Table 5 shows weak correlations of the low-effort condition with different

[4] We included the manipulation check "I was instructed to respond to all questions without effort" with response options ranging from 1 (*disagree*) to 5 (*agree*). We did not find a significant mean difference between the control ($M_C = 3.21$) and the treatment group ($M_T = 3.30$; $t(460) = .61$, $p = .54$).

careless responding indices. Although we designed the questionnaire to provoke careless responding in all participants, we were surprised that the low-effort treatment was largely ineffective.

General Discussion of Studies 1-4

We began by comparing careful and careless respondents in Studies 1 and 2, demonstrating that using Laz.R scores led to improvements in psychometric properties. In Study 3, we evaluated Laz.R against other careless responding indices, highlighting its utility as an additional tool for detecting careless respondents. However, findings also indicated that no single index fully captured all forms of careless responding. Study 4 was designed to compare Laz.R with a wide range of precautionary approaches. To achieve this, we employed a questionnaire design specifically intended to elicit a high number of careless responses, creating an optimal setting for index comparison. This approach, often called an "extreme-groups design" (Cortina & DeShon, 1998), does not reflect the typical context for organizational researchers. While the types of careless response patterns observed in Study 4 are arguably comparable to those in other studies, the proportion of careless respondents is likely much higher than what would typically be expected in standard survey research. Therefore, we caution researchers against using the results from Study 4 as a baseline or a model for survey design in their own studies. Nevertheless, a particularly noteworthy finding from Study 4 was the low to moderate correlations between most careless responding indices. To test the generalizability of this finding, we conducted a set of supplemental analyses with the Big 5 and RIASEC datasets that we used in Studies 1 and 2 (see Table 6). In addition to post-hoc statistical approaches, we report correlations with response time, as time stamps were available for both datasets. Correlations among different indicators were also low, with only high correlations between Laz.R and Longstring (Big 5: $r = 0.63$; RIASEC: $r = 0.82$), and between the four indices IRV, Mahalanobis D, G_{Poly}, and lz_{Poly}. However, note that low

IRV and high Mahalanobis D values have been suggested to indicate careless responding, making the positive correlation less intuitive. Mathematically, the positive correlation can be attributed to the fact that both the Mahalanobis distance and IRV involve calculating the deviation of each response from a mean value. Specifically, the formula for the Mahalanobis distance includes the deviation of each response from the overall mean of each item. In contrast, IRV computes the deviation from the individual's overall mean response. Thus, IRV and Mahalanobis distance increase as responses deviate from their mean values. Lastly, we did not find high correlations of post-hoc indices with response time. However, comparing response time with actual answering patterns casts doubts on the sensitivity of the two-second-per-item rule (i.e., its ability to avoid false negatives). In the two datasets, 1,672 (Big 5) and 388 (RIASEC) individuals chose the same scale anchor throughout all items. 80.3% (Big 5) and 46.4% (RIASEC) of these extreme answering patterns were flagged as careless respondents when applying the two-seconds-per-item rule (Huang et al., 2012).

Overall, results in Studies 3 and 4 showed that the precautionary measures only identified a fraction of careless respondents but failed to capture a larger number of suspicious answering patterns, especially in longer questionnaires. While this finding aligns with Barends and de Vries (2019), we extend previous research by demonstrating the usefulness of Laz.R as an additional measure to identify those suspicious answering patterns even when precautionary measures were already implemented (see Supporting Materials D for additional analysis of precautionary approaches in five other datasets). However, additional research is required to expand upon these preliminary findings and to comprehensively understand the complexities and optimal configurations of careless respondent indices. Next, we turn to the issue of how to more precisely identify Laz.R cutoffs to distinguish careful from careless respondents.

Using the Kneedle Algorithm to Set Cutoff Values for Laz.R

Previous research (e.g., Johnson, 2005; Kurtz & Parish, 2001; Meade & Craig, 2012) and our own results showed that the percentage of careless respondents seems to differ substantially between samples. This makes a rule of thumb based on a universal cutoff value for Laz.R and other post-hoc approaches less desirable. Accordingly, we offer a method to find sample-specific cutoff values for Laz.R.

Careless responding may occur in extreme forms such as when respondents completely ignore item content and answer all questions with the same scale anchor. However, our results showed that respondents often seemed to answer only some parts of the questionnaire with insufficient effort, possibly because of temporary distractions or because they lost interest during questionnaire completion. For Laz.R, we expect a group of careless respondents with only slightly higher Laz.R scores when only a few items were clicked through, up to extreme Laz.R scores for respondents who clicked through the whole questionnaire. We believe that most survey participants are for the most part careful respondents, which should be reflected in relatively similar Laz.R scores for this group. Accordingly, when Laz.R scores are sorted from high to low, there should be a "knee" in the graph, distinguishing the careless respondents with very high to slightly higher Laz.R scores from careful respondents with relatively similar Laz.R scores. To illustrate this point, we reanalyzed data from the three datasets we used in our earlier studies (i.e., Big 5, RIASEC, and CR study). We sorted the Laz.R scores of all respondents from highest to lowest as shown in Figure 1. For example, in the top left corner of the figure, we show results from the Big 5 dataset that was introduced above. This graph starts with less than 1% of respondents that have extreme Laz.R scores, followed by a steep decline, before the graph levels off after a sharp curve – the "knee" of this graph – at about 5-10%. The graphs for the three samples are similar: they start with a steep decline in Laz.R scores, followed by a knee, and afterward, scores level off. We argue that cases on the left of this knee might be considered

careless respondents. Suppose we find such a knee point in a sample. In that case, this cutoff value can guide researchers to identify cases that should be scrutinized further and possibly excluded from further analyses because of careless responding.

In computer science, the kneedle algorithm was developed to detect knee points (i.e., points where a curve flattens out; Satopää et al., 2011). The dotted lines in the panels in Figure 1 indicate knee points from the kneedle algorithm in each sample. For example, this line is at 6.4% in the Big 5 dataset graph. Thus, we might flag the 6.4% of respondents with the most extreme Laz.R scores as careless respondents and possibly remove these cases after further inspection. The kneedle cutoff for data from the CR study is 18.3%, which is in line with our observation that this dataset contained a larger number of careless respondents (see Supporting Materials E for examples of kneedle cutoffs in other datasets and for other post-hoc indices).

Best-practice Recommendations for Using Laz.R in Combination with Existing Approaches

Survey Design

When designing a survey, we recommend using scales with reverse-coded items or other means that make the simplest "clicking-through" patterns (i.e., choosing the same scale anchor), identifiable in later analyses. Without reverse-coded items, researchers cannot distinguish careless respondents from individuals who genuinely answered the constructs by consistently choosing the same scale anchor, especially in shorter questionnaires. In addition, if researchers collect primary data or rely on secondary data including precautionary measures, we suggest combining Laz.R with precautionary measures, which is consistent with existing recommendations (e.g., DeSimone et al., 2015; Goldammer et al., 2020; Kam & Meyer, 2015). We have briefly introduced the most common precautionary measures, but a detailed analysis of existing options goes beyond the scope of this study. Note, however, that we found a larger

number of suspicious cases not flagged by the respective precautionary measure in the three datasets that we used to compare post-hoc and precautionary measures. Precautionary measures might even lose some of their power when respondents get paid for survey participation. Some respondents may engage in careless answering or utilize bots while deliberately attempting to pass precautionary items to secure their compensation. Thus, if researchers are concerned that bots generated their survey responses, they may consider incorporating specific bot detection measures (Xu et al., 2022).

Data Analysis

Based on our results, we first recommend the use of the Laz.R index to identify *non-random patterned* responses, like straightlining and seesaw responding. Our analyses have shown that the Laz.R index captures forms of patterned responses that other indices overlook. Specifically, for seesaw response patterns such as 4-5-4-5 or 1-2-3-4-5-4-3-2-1, the Laz.R index demonstrates a superior ability to detect these consistently patterned response structures compared to other indices.

Researchers should include all items with a similar number of answering options from their survey instrument (e.g., all items from 5-point scales). Note that a lower number of items might produce more false positives. For example, if only four items are used for index computation, the pattern 5-5-5-5 results in extreme scores for Laz.R, longstring, IRV, and Mahalanobis D, although it might reflect true answering behavior. We thus suggest using twenty or more items for index computation. If desired, missing data can be added as an additional category to the scale anchors. We further recommend conducting a combined analysis of Laz.R results to identify non-random patterned responses, alongside utilizing Psychometric/Semantic Synonyms or a person-fit statistic (e.g., G_{Poly} and lz_{Poly}) to detect *random responses*. Precautionary measures should also be considered if available. Researchers should report their

approach and the number of deleted cases in publications.

Moreover, we advocate employing the kneedle algorithm to determine cutoff values for post-hoc approaches. Thus, for Laz.R, we suggest that all cases that have a score above the kneedle cutoff should be inspected and removed from further analyses, except if there is evidence that the respective response pattern emerged from careful responding.

Finally, for easy and accessible use, we implemented our approach to detecting careless responding in an interactive web application, using the *shiny* package in R (Chang et al., 2015). The R Shiny app is available at https://hrmmannheim.shinyapps.io/ShinyCR_App/ and guides users through a step-by-step process from uploading the data to initiating the computations of post-hoc indices and cutoff values. With this app, we make the computation of Laz.R and other common post-hoc approaches readily available for fellow researchers and practitioners. Figure 2 summarizes our recommendations.

Conclusions

Laz.R makes explicit that some careless respondents take a low-effort route and "click through" the survey, disregarding the content of specific items. We analyzed three datasets and provided evidence that the use of Laz.R improves psychometric properties and the accuracy of substantive conclusions. For example, in Study 1, the mean correlation among the theoretically independent Big 5 personality dimensions was r = .13 for careful respondents (i.e., those with high Laz.R scores) and r = .37 for careless respondents (i.e., those with low Laz.R scores). In Study 2, correlations of the three theoretically opposite subdimensions of the RIASEC questionnaire were r = .04/-.01/-.01 for careful respondents, but – contrary to theory – strongly positive for careless respondents with r = .47/.51/.60. These results indicate that conclusions across various OB/HRM domains are biased by overly positive relationships if researchers fail to detect careless respondents. Based on its consistently superior performance across all datasets, we

recommend the use of Laz.R for detecting patterned responses. Additionally, we suggest that researchers employ Psychometric/Semantic Synonyms and person-fit statistics to identify random responses. For index computation, we encourage the use of the user-friendly R Shiny app, which requires no R knowledge, combined with precautionary approaches, to minimize the detrimental effects of careless respondents on substantive conclusions.

References

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301. https://doi.org/10.1177/1094428112470848

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*, 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

Baer, R. A., Ballenger, J., Berry, D. T., & Wetter, M. W. (1997). Detection of random responding on the MMPI--A. *Journal of Personality Assessment*, *68*(1), 139–151. https://doi.org/10.1207/s15327752jpa6801_11

Barends, A. J., & de Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences*, *143*, 84–89. https://doi.org/10.1016/j.paid.2019.02.015

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology: Interdisciplinary and Applied*, *123*(1), 101–103. https://doi.org/10.1080/00223980.1989.10542966

Bosco, F. A., Aguinis, H. Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Bowling, N.A., Huang, J.L., Brower, C.K., & Bragg, C.B. (2023). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*, *26*(2), 323–352. https://doi.org/10.1177/10944281211056520

Bowling, N. A., Wagner, S. H., & Beehr, T. A. (2018). The *Facet Satisfaction Scale*: An effective affective measure of job satisfaction facets. *Journal of Business and Psychology*, *33*, 383–403. https://doi.org/10.1007/s10869-017-9499-4

Burns, G. N., Morris, M. B., Periard, D. A., LaHuis, D., Flannery, N. M., Carretta, T. R., & Roebke, M. (2017). Criterion-related validity of a Big Five general factor of personality from the TIPI to the IPIP. *International Journal of Selection and Assessment*, *25*(3), 213–222. https://doi.org/10.1111/ijsa.12174

Cable, D. M., & DeRue, D. S. (2002). The convergent and discriminant validity of subjective fit perceptions. *Journal of Applied Psychology*, *87*(5), 875–884. https://doi.org/10.1037/0021-9010.87.5.875

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2015). Package 'shiny'.

Cortina, J. M., & DeShon, R. P. (1998). Determining relative importance of predictors with the observational design. *Journal of Applied Psychology*, *83*(5), 798–804. https://doi.org/10.1037/0021-9010.83.5.798

Costa, P.T., Jr., & McCrae, R.R. (2008). The revised NEO personality inventory (NEO-PI-R). In D.H. Saklofske (Ed.), *The SAGE handbook of personality theory and assessment: Volume 2 — Personality measurement and testing* (pp. 179–198). Sage Publications, Inc. https://doi.org/10.4135/9781849200479.n9

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

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, *67*(2), 309–338. https://doi.org/10.1111/apps.12117

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171–181. https://doi.org/10.1002/job.1962

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

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

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 and Psychology*, *33*, 105–121. https://doi.org/10.1007/s10869-016-9479-0

Ehrhart, K. H., Roesch, S. C., Ehrhart, M. G., & Kilian, B. (2008). A test of the factor structure equivalence of the 50-item IPIP Five-factor model measure across gender and ethnic groups. *Journal of Personality Assessment*, *90*(5), 507–516. https://doi.org/10.1080/00223890802248869

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

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, *75*, 47–52. https://doi.org/10.1016/j.paid.2014.11.017

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), 101384. https://doi.org/10.1016/j.leaqua.2020.101384

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Goldberg, L.R., & Kilkowski, J.M. (1985). The prediction of semantic consistency in self descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology*, *48(1)*, 82–98. https://doi.org/10.1037/0022-3514.48.1.82

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. *Academy of Management Journal*, *33*(1), 64–86. https://doi.org/10.2307/256352

Hellgren, J., Sverke, M., & Isaksson, K. (1999). A two-dimensional approach to job insecurity: Consequences for employee attitudes and well-being. *European Journal of Work and Organizational Psychology*, *8*(2), 179–195. https://doi.org/10.1080/135943299398311

Hill, N. S., Aguinis, H., Drewry, J. M., Patnaik, S., & Griffin, J. 2022. Using macro archival databases to expand theory in micro research. *Journal of Management Studies, 59*(3), 627–659. https://doi.org/10.1111/joms.12764

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Psychological Assessment Resources.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015a). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299–311. https://doi.org/10.1007/s10869-014-9357-6

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort respond to surveys. *Journal of Business and Psychology*, *27*, 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., & DeSimone, J. A. (2021). Insufficient effort responding as a potential confound between survey measures and objective tests. *Journal of Business and Psychology*, *36*(5), 807–828. https://doi.org/10.1007/s10869-020-09707-2

Huang, J. L., Liu, M., & Bowling, N. A. (2015b). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Hurtado Rúa, S. M., Stead, G. B., & Poklar, A. E. (2019). Five-factor personality traits and RIASEC interest types: A multivariate meta-analysis. *Journal of Career Assessment*, *27*(3), 527–543. https://doi.org/10.1177/1069072718780447

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, *52*(3), 621–652. https://doi.org/10.1111/j.1744-6570.1999.tb00174.x

Kam, C. C. S. (2019). Careless responding threatens factorial analytic results and construct validity of personality measure. *Frontiers in Psychology*, *10*, 1258. https://doi.org/10.3389/fpsyg.2019.01258

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541. https://doi.org/10.1177/1094428115571894

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment*, *76*(2), 315–332. https://doi.org/10.1207/S15327752JPA7602_12

Liao, H. Y., Armstrong, P. I., & Rounds, J. (2008). Development and initial validation of public domain Basic Interest Markers. *Journal of Vocational Behavior*, *73*(1), 159–183. https://doi.org/10.1016/j.jvb.2007.12.002

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*(1), 49‑55.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

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. https://doi.org/10.1016/j.paid.2014.08.021

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

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*(2), 111–120. http://dx.doi.org/10.1177/014662169401800202

Motowidlo, S. J., Packard, J. S., & Manning, M. R. (1986). Occupational stress: Its causes and consequences for job performance. *Journal of Applied Psychology*, *71*(4), 618–629. https://doi.org/10.1037/0021-9010.71.4.618

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

R Core Team. (2024). R: A language and environment for statistical computing. https://www.r-project.org/.

Satopää, V., Albrecht, J., Irwin, D., & Raghavan, B. (2011). Finding a" kneedle" in a haystack: Detecting knee points in system behavior. *31st International Conference on Distributed Computing Systems Workshops* (pp. 166–171). IEEE. https://doi.org/10.1109/ICDCSW.2011.20

Song, Q. C., Shin, H. J., Tang, C., Hanna, A., & Behrend, T. (2024). Investigating machine learning's capacity to enhance the prediction of career choices. *Personnel Psychology*, *77*(2), 295–319. https://doi.org/10.1111/peps.12529

Spreitzer, G. M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal*, *38*(5), 1442–1465. https://doi.org/10.2307/256865

Stark, S., Chernyshenko, O. S., Nye, C., D., Drasgow, F., & White, L. A. (2017). *Moderators of the Tailored Adaptive Personality Assessment System Validity (Technical Report 1357)*. Fort Belvoir, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (Available from the on-line Defense Technical Information Center)

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*(6), 859–884. https://doi.org/10.1037/a0017364

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

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*(3), 267–298. https://doi.org/10.1177/0022002182013003001

Van Iddekinge, C. H., Ferris, G. R., & Heffner, T. S. (2009). Test of a multistage model of distal and proximal antecedents of leader performance. *Personnel Psychology*, *62*(3), 463–495. https://doi.org/10.1111/j.1744-6570.2009.01145.x

Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology*, *96*(6), 1167–1194. https://doi.org/10.1037/a0024343

Ward, M. K., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*, 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Xu, Y., Pace, S., Kim, J., Iachini, A., King, L. B., Harrison, T., DeHart, D., Levkoff, S. E., Browne, T. A., Lewis, A. A., Kunz, G. M., Reitmeier, M., Utter, R. K., & Simone, M. (2022). Threats to online surveys: Recognizing, detecting, and preventing survey bots. *Social Work Research*, *46*(4), 343–350. https://doi.org/10.1093/swr/svac023

Yentes R.D., & Wilhelm, F. (2021). *careless: Procedures for computing indices of careless responding*. R package version 1.2.1.

Table 1

Examples of Patterned Responses and corresponding Laz.R scores with a varying number of items

		Laz.R index				
Number of scale anchors	Straightlining with the same scale anchor	First 10 items	First 20 items	First 30 items	First 40 items	All 50 items
any	1111111111 1111111111 1111111111 1111111111 1111111111 2222222222 2222222222 2222222222 2222222222 2222222222 … 7777777777 7777777777 7777777777 7777777777 7777777777	1.00	1.00	1.00	1.00	1.00
	Straightlining with moving scale anchor					
5	1111111111 2222222222 3333333333 4444444444 5555555555	1.00	0.91	0.88	0.86	0.85
5	1112223334 4455511122 2333444555 1112223334 4455511122	0.56	0.57	0.58	0.56	0.56
7	1111122222 3333344444 5555566666 7777711111 2222233333	0.82	0.75	0.72	0.71	0.70
7	1112223334 4455566677 7111222333 4445556667 7711122233	0.56	0.58	0.57	0.56	0.56
	Seesawing					
any	5151515151 5151515151 5151515151 5151515151 5151515151	1.00	1.00	1.00	1.00	1.00
any	4545454545 4545454545 4545454545 4545454545 4545454545	1.00	1.00	1.00	1.00	1.00
5	1234543212 3454321234 5432123454 3212345432 1234543212	0.67	0.64	0.64	0.63	0.63
7	1234567654 3212345676 5432123456 7654321234 5676543212	0.78	0.65	0.60	0.59	0.59
	Random answers (uniform distribution)					
5	4521421511 3443542155 1124433442 2545524215 5414515435	0.56	0.41	0.33	0.29	0.27
5	4555513133 4254524451 4335514135 3224333125 4225415135	0.72	0.44	0.35	0.30	0.28
7	1127262532 1364576441 4726745466 4475176727 5621474615	0.67	0.40	0.30	0.27	0.24
7	5352132231 1143266411 4117741755 3212152612 3136144625	0.44	0.41	0.37	0.34	0.27

Table 2

Study 1 Results: Reliability estimates (Cronbach's alpha) and Zero-order Correlations between Big 5 Personality Dimensions for Careful and Careless Respondents and Comparison with Results by Burns et al. (2017) and Ehrhart et al. (2008)

	Items[a]		**α**	**E**	**A**	**C**	**N**
Extraversion (E)	10 (5)	Careful	0.90	-			
		Careless	0.88	-			
		Burns et al. (2017)	0.90				
		Ehrhart et al. (2008)	0.89				
Agreeableness (A)	10 (4)	Careful	0.84	.29	-		
		Careless	0.84	.43	-		
		Burns et al. (2017)	0.79	.19			
		Ehrhart et al. (2008)	0.78	.32			
Conscientiousness (C)	10 (4)	Careful	0.82	.04	.13	-	
		Careless	0.81	.36	.41	-	
		Burns et al. (2017)	0.79	.09	.24		
		Ehrhart et al. (2008)	0.81	.03	.16		
Neuroticism[b] (N)	10 (2)	Careful	0.87	-.20	-.03	-.22	-
		Careless	0.89	-.50	-.25	-.44	-
		Burns et al. (2017)	0.87	-.25	-.13	-.29	
		Ehrhart et al. (2008)	0.86	-.21	-.07	-.12	
Openness (O)	10 (3)	Careful	0.80	.15	.09	.04	-.08
		Careless	0.83	.29	.45	.41	-.18
		Burns et al. (2017)	0.80	.25	.30	.25	-.15
		Ehrhart et al. (2008)	0.78	.33	.26	.14	-.21

[a] Number of reverse-coded items in parentheses

[b] Labeled *emotional stability* in Burns et al. (2017) and Ehrhart et al. (2008); hence, correlations with neuroticism were reversed.

Table 3

Study 2 Results: Reliability estimates (Cronbach's alpha) and Zero-order Correlations between Holland Occupational Themes for Careful and Careless Respondents and Comparison with Results by Hurtado Rúa et al. (2019)

	Items[a]		**α**	**R**	**I**	**A**	**S**	**E**	**C**	**Gender differences**
Realistic (R)	8 (0)	Careful	0.87	-						0.80
		Careless	0.96	-						0.59
	-	Hurtado Rúa et al. (2019)		-						0.84
Investigative (I)	8 (0)	Careful	0.89	.26	-					0.08
		Careless	0.94	.55	-					-0.01
	-	Hurtado Rúa et al. (2019)		.41	-					0.26
Artistic (A)	8 (0)	Careful	0.85	.14	.28	-				-0.04
		Careless	0.93	.61	.51	-				0.14
	-	Hurtado Rúa et al. (2019)		.07	.24	-				-0.35
Social (S)	8 (0)	Careful	0.83	.04	.16	.29	-			-0.45
		Careless	0.92	.47	.49	.51	-			-0.39
	-	Hurtado Rúa et al. (2019)		-.04	.19	.41	-			-0.68
Enterprising (E)	8 (0)	Careful	0.82	.28	-.01	.25	.36	-		0.01
		Careless	0.94	.75	.51	.67	.59	-		0.18
	-	Hurtado Rúa et al. (2019)		.17	.17	.27	.40	-		0.04
Conventional (C)	8 (0)	Careful	0.89	.45	.05	-.04	.14	.48	-	0.15
		Careless	0.96	.78	.53	.60	.55	.84	-	0.25
	-	*Hurtado Rúa et al. (2019)*		.16	.19	.18	.22	.51	-	*-0.33*

[a] Number of reverse-coded items in parentheses

Table 4

Study 3 Results: Percentile Ranks and Values for Careless Responding Indices: Laz.R., Longstring, Mahalanobis Distance (MD), Intra-individual Response Variability (IRV), Psychometric/semantic Synonyms (PsychSyn) and Antonyms (PsychAnt), G_{Poly}, and lz_{Poly}

Questionnaire structure for the constructs Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness (O) (reverse-coded items are underlined): **EACNO EACNO EACNO EACNO EACNO EACNO EACNO EACNO EACNO EACNO**											
	Laz.R	Longstr.	MD	IRV_{low}	PsychSyn			PsychAnt		G_{Poly}	lz_{Poly}
					r_{40} 56 pairs	r_{50} 24 p.	r_{60} 5 p.	r_{-40} 29 p.	r_{-50} 11 p.		
Examples of Response Patterns	**Percentile ranks**										
Straightlining – Lowest scale anchor:											
11111 11111 11111 11111 11111 11111 11111 11111 11111 11111	1	1	2	1	*NA*	*NA*	*NA*	*NA*	*NA*	1	1
Straightlining – Moderately high-scale anchor:											
44444 44444 44444 44444 44444 44444 44444 44444 44444 44444	1	1	11	1	*NA*	*NA*	*NA*	*NA*	*NA*	32	49
Straightlining – Moving scale anchor:											
11111 11112 22222 22233 33333 33333 34444 44444 44444 55555	1	1	4	44	53	38	7	1	1	2	4
Seesaw – Diagonal-lining:											
12345 43212 34543 21234 54321 23454 32123 45432 12345 43212	1	100	3	41	1	2	16	8	6	5	7
Seesaw – Extreme variance:											
51515 15151 51515 15151 51515 15151 51515 15151 51515 15151	1	100	2	100	44	21	1	75	100	1	2
Seesaw – Medium variance:											
45454 54545 45454 54545 45454 54545 45454 54545 45454 54545	1	100	17	1	44	21	1	75	100	1	1
Random responses	**Mean percentile ranks (Sample plus N = 1,000 simulated random cases)**										
Random responses – uniform distribution (20/20/20/20/20)	84.0	60.4	1.7	66.4	3.2	3.7	13.1	6.7	9.8	2.1	2.0
Random responses – central tendency (10/20/40/20/10)	46.7	40.6	10.7	20.8	3.3	3.9	13.6	6.9	10.6	12.4	12.6
Random responses – skewed distribution (5/10/15/30/40)	33.4	36.0	3.4	30.8	3.8	4.9	15.1	6.9	10.0	1.8	1.6

Table 5

Study 4 results: Descriptives and Zero-order Correlations between Careless Responding Indices

	Post-hoc procedures									*Precautionary Approaches*				
	Laz.R	Longstr.	MD	IRV[a]	PsychSyn[a]			G_{Poly}[a,b]	lz_{Poly}[b]	Failed resp. time	Self-report[a]	Infr. items	Failed instr. resp.	Cont. recog.
					r_{40}	r_{50}	r_{60}							
Min	0.25	1	1.50	0.00	-0.05	-0.17	-1.00	0.00	-4.21	0	19	3	0	0
Max	1.00	103	270.1	1.81	0.78	0.92	1.00	0.61	2.31	4	45	15	3	10
Mean	0.49	6.38	102.8	0.92	0.11	0.17	0.06	0.18	0.24	0.61	28.7	9.6	0.25	6.0
SD	0.16	10.39	42.47	0.28	0.17	0.21	0.46	0.12	1.06	0.73	5.14	3.66	0.56	2.55
Laz.R	-													
Longstr.	.42	-												
MD	-.62	-.18	-											
IRV[a]	-.66	-.26	.79	-										
Psych Syn[a] – r_{40}	-.37	.00	.15	.68	-									
Psych Syn[a] – r_{50}	-.24	.07	.12	.62	.91	-								
Psych Syn[a] – r_{60}	-.10	.10	.10	.24	.26	.31	-							
G_{Poly}[a,b]	-.28	.15	.55	.52	.26	.25	.02	-						
lz_{poly}[b]	.40	.02	-.77	-.81	-.36	-.35	-.13	-.64	-					
Failed resp.time	.26	.28	-.12	-.27	-.33	-.27	-.14	.12	.06	-				
Self-report[a]	-.27	-.15	.13	.45	.59	.53	.26	.12	-.20	-.30	-			
Infr. items	.47	.25	-.33	-.56	-.57	-.47	-.26	-.26	.23	.38	-.55	-		
Failed instr. resp.	.28	.29	.00	-.22	-.11	-.07	-.04	.06	.06	.15	-.13	.14	-	
Cont. recog.	.26	.14	-.07	-.38	-.57	-.50	-.18	-.02	.13	.39	-.51	.50	.18	-
Treatment	.06	.04	-.06	-.08	-.04	-.07	-.05	-.03	.05	.00	-.11	.09	.09	.02

Notes. N = 465; Longstr. = Longstring index; MD = Mahalanobis Distance; IRV = Item Response Variability; PsychSyn = Psychological Synonyms; PsychAnt = Psychological Antonyms; Failed resp. time = Number of pages with failed response time (two-second-per-item rule); Infr. Items = Infrequency items; Failed instr. resp. = Number of items with failed instructed response; Cont. recog. = Number of items with correctly recognized content.

[a] Low index values signal careless responding; hence, negative correlation coefficients indicate coherence with other indices.

[b] Computed with the 50 IPIP items.

Table 6

Zero-order Correlations between Careless Responding Indices for the Big 5 (below diagonal) and the RIASEC dataset (above diagonal)

	Laz.R	Longstr.	MD	IRV[a]	PsychSyn[a]			PsychAnt		G_{Poly}[a]	lz_{Poly}	Failed resp. time		
					r_{40}	r_{50}	r_{60}	r_{-40}	r_{-50}			1sec	2sec	3sec
Laz.R	-	.82	-.16	-.21	.11	.07	.08	-	-	-.22	.09	.05	.20	.25
Longstring	.63	-	-.12	-.17	.16	.16	.14	-	-	-.18	.07	.05	.20	.22
Mahalanobis D	.02	.07	-	.64	-.36	-.35	-.31	-	-	.93	-.90	-.01	-.05	-.09
IRV[a]	-.23	-.28	.52	-	.35	.34	.28	-	-	.62	-.66	-.05	-.09	-.02
Psych Syn[a] – r_{40}	-.03	-.19	-.33	.31	-	.91	.70	-	-	-.39	.28	-.01	-.01	.09
Psych Syn[a] – r_{50}	-.05	-.18	-.33	.28	.85	-	.78	-	-	-.33	.25	-.02	-.02	.08
Psych Syn[a] – r_{60}	-.07	-.12	-.16	.16	.37	.45	-	-	-	-.26	.20	-.02	-.01	.07
Psych Ant – r_{-40}	.07	.16	.33	-.28	-.66	-.61	-.25	-	-	-	-	-	-	-
Psych Ant – r_{-50}	.08	.14	.29	-.23	-.55	-.52	-.26	.82	-	-	-	-	-	-
G_{Poly}[a]	.08	.18	.94	.42	-.36	-.32	-.12	.35	.29	-	-.93	-.02	-.06	-.09
lz_{Poly}	-.05	-.15	-.93	-.53	.27	.26	.10	-.26	-.23	-.96	-	.01	.03	.04
Failed resp. time – 1sec	.26	.40	.03	-.12	-.04	-.04	-.03	.04	.04	.09	-.08	-	.24	.08
Failed resp. time – 2sec	.26	.36	-.11	-.11	-.06	-.03	-.03	.03	.03	.07	-.06	.36	-	.32
Failed resp. time – 3sec	.11	.12	-.08	-.03	.07	.07	.02	-.08	-.07	-.07	.06	.09	.25	-

Notes. Longstr. = Longstring index; MD = Mahalanobis Distance; IRV = Item Response Variability; PsychSyn = Psychological Synonyms; PsychAnt = Psychological Antonyms (no pairs of items with $r < -.40$ / $r < -.50$ to compute PsychAnt in the RIASEC dataset); Failed resp. time = Number of pages with failed response time (1/2/3-seconds-per-item rule).

[a] Low index values signal careless responding; hence, negative correlation coefficients indicate coherence with other indices.

Figure 1

Kneedle Cutoff Values to Distinguish Careful from Careless Respondents across Samples using Laz.R

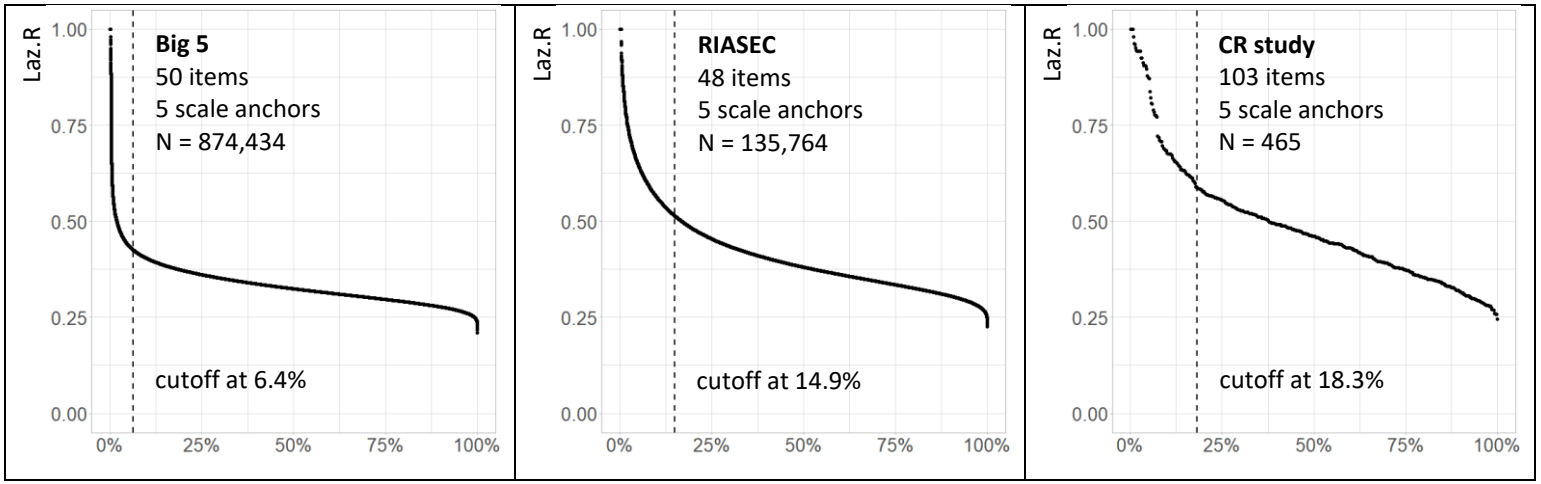

Figure 2

Summary of Recommendations and Tools for Identifying Different Types of Careless Responding