June 2022

Prevalence and Predictors of Careless Responding in Experience Sampling Research

Alexander J. Denison
University of South Florida

Prevalence and Predictors of Careless Responding in Experience Sampling Research

by

Alexander J. Denison

A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Arts
with a concentration in Industrial-Organizational Psychology
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Kelsey L. Merlo, Ph.D.
Brenton M. Wiernik, Ph.D.
Fallon R. Goodman, Ph.D.

Date of Approval:
May 6, 2022

Keywords: Insufficient Effort Responding, Aberrant Responding, ESM, Detection Methods

Table of Contents

List of Tables

List of Figures

Abstract

In the current study we examine the prevalence and several predictors of careless responding to an experience sampling (ESM) study. While careless responding has been noted as a potential problem in ESM research, few studies have examined the prevalence of this behavior (Beal, 2015; Berkel et al., 2017; Eisele et al., 2020; Gabriel et al., 2019; Jaso et al., 2021). Using statistical methods of careless response classification, we derive cut scores from data simulation and graphical examination of item correlations, and flag 44.98% of response episodes as careless. A majority of these flagged episodes were the product of overly consistent response patterns, such as long strings of identical responses or low variance response patterns. Further analyses revealed that careless responding increased significantly over time and was associated with several personality variables. Taken together, these results indicate that careless responding is a serious issue in ESM studies and is related to both study-level and individual-level factors.

Chapter 1: Introduction

During the past decade, advances in technology have spurred the development or improvement of a variety of assessment methods in psychological science. One method that has seen such advancement is the Experience Sampling Methodology (ESM), which is a broad group of methods that share three common features: capturing an individual's experience as closely to their natural form as possible, focusing on concrete or immediate experiences over more abstract or distal ones, and assessing a range of experiences that accurately reflect an individual's daily life (Beal, 2015). Generally, an individual's experiences are sampled one or more times throughout the day over the course of several days or weeks. By assessing responses at different timepoints ESM provides rich data at the individual level and allows researchers to assess variability or growth in psychological characteristics over short periods of time (Beal, 2015).

While experience sampling methods initially utilized non-technological methods, such as daily diaries, the wide availability of smartphones has offered a convenient and affordable way for researchers to obtain ESM data. A common approach to smartphone data collection is to use an app to prompt people to respond to a survey on their phone, either once a day, periodically throughout the day, or in response to some critical event (Berkel et al., 2017). These response "episodes" can then be analyzed to examine individual variability over time. However, accurate insights about changes in self-reported psychological constructs can only be drawn if individuals respond to these surveys in a thoughtful and honest way. It is well accepted that participants do not always do this, and may instead engage in a variety of aberrant response patterns (Curran, 2016; Karabatsos, 2003; van de Mortel, 2008; Weijters et al., 2010).

One aberrant response pattern that may be especially likely in ESM is careless responding. Participants are considered careless when they do not put forth the required effort to respond to a survey thoughtfully and truthfully, that is, their responses do not reflect their latent standing on the measured construct. Careless responding is differentiated from other response styles, such as faking, in that it is content non-responsive behavior; participants do *not* pay attention to the items they are responding to or attempt to manipulate their results in some way (Curran, 2016).

Careless responding should be of special concern in ESM studies due to three factors: (1) participants are often asked to respond to the same self-report questions multiple times throughout the study, so may stop paying attention to the content of the specific items as the study progresses. (2) individuals are going about their daily lives as data is being collected, meaning they may respond to questions while they are distracted by the external environment. (3) individuals may become bored of responding to the same questions as the study progresses and exert less effort when providing their responses as a result.

Current Literature on Careless Responding and ESM

A number of authors have discussed careless responding as a potential problem in ESM when examining the state of the literature (c.f. Beal, 2015; Berkel et al., 2017; Gabriel et al., 2019); however, only two recently published studies have investigated this topic (Eisele et al., 2020; Jaso et al., 2021). This lack of work is troubling, as incorrect conclusions can be drawn from statistical tests when careless respondents make up as little as 5% of a studies total data (Huang et al., 2015). Data from careless responses can generate spurious relationships between variables, increasing Type I error (DeSimone et al., 2018; Huang et al., 2015), or obscure meaningful relationships by adding random noise, increasing Type II error (Huang et al., 2012; Kam & Meyer, 2015; McGonagle et al., 2016; Schneider et al., 2018).

Of the two studies on careless responding in ESM, Eisele et al., used retrospective self-reports of careless responding to test if different sampling frequencies and questionnaire lengths increased careless response rates, but did not provide an estimate of the prevalence of careless response rates (2020). Jaso et al., (2021) used statistical classification, discussed below, to identify careless respondents and found that 60% of participants had at least one response episode that was classified as careless, and 5.46% had at least 50% of their response episodes throughout the study classified as carless using the most conservative cutoff score (38.57% using a more liberal cutoff). This provides preliminary evidence that careless responding occurs in ESM studies at a non-trivial frequency.

The current study seeks to add to the growing literature on careless responding in ESM data by exploring the frequency and characteristics of careless responding in an ESM study using cellphone assessment. Specifically, the following five questions will be examined. (1) What is the prevalence of careless responding at baseline and during the ESM study and do these frequencies algin with the estimates reported by Jaso et al.? (2) Do careless respondents change the relationship among daily item correlations? (3) Is there variability in who responds carelessly throughout the study, or are the same people repeatedly careless? (4) Are participants more likely to respond carelessly at each episode as their time of participation in the study increases? (5) Do personality scores at baseline predict careless responding during the ESM portion of the study? Conscientiousness and agreeableness should negatively correlate with careless response rates; however, it is unclear how or whether other individual differences will. In answering these questions, this paper also provides suggestions for how researchers might deal with careless responding when participants are assessed at multiple timepoints.

This study expands upon the work of Jaso et al., (2021) by examining additional metrics for detecting carelessness. Additional analyses are also conducted to assess how time and theoretically relevant personality variables influences careless response rates. Finally, this study employs a working sample that received competitive monetary compensation for their participation, compared to the student sample of Jaso et al., who received course credit.

Detecting Careless Responding

One potential hurdle that has prevented more work in this area is the challenge of determining who is a careless respondent and who is not. Multiple methods have been developed for assessing careless responses to traditional, single-timepoint surveys, however moving to the rapid, multi-timepoint sampling procedure utilized by ESM studies causes some challenges for these traditional methods of screening.

Two general methods have been used to detect careless responses to surveys, one method involves adding content to the survey itself to detect careless respondents and the other method involves statistical analysis to detect aberrant response patterns to surveys that are assumed to be careless (Curran, 2016). Methods that involve adding survey content usually utilize questions that instruct participants to select a specific response option (e.g., select "5" on this question), ask questions that have a clear incorrect or impossible answer (e.g., I am paid biweekly by leprechauns), or directly ask participants if they were careless (e.g., Did you respond carelessly to these questions? You will not be penalized for your answer.) (Curran, 2016; Curran & Hauser, 2019). While this method is feasible for traditional studies that collect responses to many questions, it is less feasible to incorporate into the short surveys used in ESM. This is because ESM studies typically have very tight constraints on the number of items in the study, and multiple instructed response items are needed to accurately gauge careless responding (Curran & Hauser, 2019). Furthermore, as

participants are repeatedly asked the same questions in ESM studies, they may realize the purpose of these items and modify their responses to not seem careless.

Conversely, indirect measures of careless responding can be calculated on most Likert-type or response slider self-report measures that assess psychological constructs, meaning that researchers do not need to modify their study to detect careless responding. Indirect measures use different statistical approaches to identify response patterns that are extremely unlikely for a conscientious respondent. For example, given five positively worded items that measure extraversion on a five-point Likert scale, the response sets of [1, 5, 4, 2, 1] and [1, 1, 1, 5, 5] are both unlikely for a conscientious participant. Assuming the scale measures a unidimensional construct, participant responses should be largely consistent with each other.

Statistical methods of careless response detection are generally designed to flag two types of unlikely response patterns, overly inconsistent responses or overly consistent responses (Curran, 2016). Inconsistent responding (also called random responding) is characterized by a large degree of variability within a participant's response to a scale. Consistent responding is characterized by a long string of the same response option (e.g. selecting 1 for every question) or some pattern within their responses (e.g. 1, 2, 1, 2 …).

Because of the different response patterns that make up careless responding, researchers have developed different techniques that are designed to detect these patterns. By using multiple techniques to screen for careless respondents' researchers can cover the weaknesses that any single method may have. Employing multiple methods can also allow researchers to set more conservative cut-off values to minimize false-positive rates. For example, if method A is great at detecting purely random responses and decent at detecting patterned responses, a researcher would either have to use a more liberal cutoff to detect patterned responses (and risk erroneously flagging non-careless

respondents) or accept that a subset of patterned respondents would go undetected. However, if method B is excellent at detecting patterned responding then using both of these methods together would allow the researcher to detect more careless respondents while minimizing erroneous flagging of non-careless respondents.

While methods that detect the same response pattern will correlate with each other there can be substantial variability between detection metrics. For example, one inconsistent careless respondent might always select a different response option on consecutive questions, whereas another inconsistent respondent may often select the same response option on consecutive responses while still remaining inconsistent in general. Finally, these methods have been developed using traditional single-timepoint assessment and it is unclear whether some of these metrics generalize to detecting careless respondents at multiple timepoints. Several methods are discussed below that were chosen to provide coverage of a variety of possible response patterns while still theoretically performing well when flagging careless respondents at multiple response episodes (Curran, 2016; Curran & Denison, 2019).

Careless Response Detection Metrics

Response time is a basic, but effective way to identify some careless respondents (Curran, 2016). This detection method does not focus on a specific response pattern, but capitalizes on the basic motivation behind careless responding, to get through a study as fast as possible. Since careless respondents do not pay attention to the items they are responding to, they will be able to complete the survey faster than someone reading every item. This means that some careless respondents will have response times that are impossible for a thoughtful respondent. For instance, if someone responds to 50 items in only 15 seconds it is safe to say that it is impossible for them to have read the questions. However, beyond impossible response-strings it is difficult to determine if someone is truly careless or is simply an especially fast respondent. Therefore, if a careless respondent has a time that is still

plausible or does something that artificially increases their response time (e.g. waiting on a page for a while to not seem suspicious) they will avoid detection.

Consistent/Inconsistent Responding Metrics

Longstring analysis

Captures overly consistent responding by calculating the longest consecutive string of the same response option for each participant (Johnson, 2005). Scores can range from 1 (no identical responses), to n, where n is the total number of items on a scale (a score of n would reflect a participant choosing the same response for every item). In the example below, this individual would have a longstring value of four, since they selected the option "2" four times in a row. Note that "1" is the most common response option (picked five times), but it was only selected consecutively a maximum of three times.

[1, 3, 1, 2, 2, 2, 2, 1, 1, 1]

Since this method only captures continuous strings of identical responses, a participant who changes their response to a new point on the scale at any point would start a new string count. A scree-like plot can be used to visually judge where sudden decreases in the probability of a given long-string score occur, indicating respondents below that point are likely responding carelessly or researchers can set a cut score that is equal to the maximum number of items that measure a given construct (Johnson, 2005).

Inter-item standard deviation

Intra-individual response variability (IRV) (Dunn et al., 2018) and inter-item standard deviation (ISD) (Marjanovic et al., 2015) are two metrics which were proposed separately, but are more or less identical in calculation. Marjanovic originally proposed this metric as a way to detect inconsistent responding by examining the standard deviation of a participant's responses. If a participant is responding to a unidimensional scale, logically, there should be a large degree of

consistency in their responses and a low standard deviation. Thus, large standard deviations may indicate that a participant is careless.

Conversely, Dunn et al. propose that an overly small standard deviation across constructs could indicate that participants are responding carelessly with an overly consistent response pattern. They propose that this metric may be more sensitive to overly consistent response patterns than longstring because it is not "fooled" by participants changing their response patterns. Consider a participant who responds with the following response pattern across a scale measuring the five factors of personality:

[1, 2, 1, 2, 1, 2, 1, 2, 1, 2]

This response pattern would have a longstring score of one, since they never select the same option consecutively. However, this respondent would have a much lower standard deviation than a conscientious respondent since we would expect a conscientious respondent to have some variability across the five personality traits. Thus, both high and low standard deviations could indicate careless responding, depending on whether the standard deviation is calculated within or between unidimensional constructs.

Hybrid Metrics

Psychometric synonyms and antonyms

These methods detect either overly consistent or overly inconsistent responding. These are pairs of items that are identified as having highly positive (synonyms) or highly negative (antonyms) correlations because they measure the same construct or are pairs of reverse-worded and positively-worded items (Johnson, 2005). Once one or multiple pairs of these items are identified, a correlation between these pairs can be calculated for each individual in a dataset. High positive correlations for synonyms and high negative correlations for antonyms, relative to the average correlation in the

sample, reflect thoughtful response, near-zero or oppositely signed scores on each metric reflect carelessness.

While psychometric synonyms will mostly detect inconsistent responding, psychometric antonyms can detect both overly consistent and inconsistent responding. For example, if reverse worded items are present in a scale, when an individual responds with a long-string or pattern of responses they may select the same option for both a reverse-worded and non-reverse-worded item when responses to these items should be opposite each other. Similarly, someone responding completely at random could choose similar options for these two items. The major downside of these two approaches is that a scale must contain items that are built to have pairs of items that share these strong positive or negative relationships.

Sample Outlier Analyses

Person total correlation

Generally detects overly inconsistent responding and is based on the method of item-total correlation (Curran, 2016; Donlon & Fischer, 1968). In item-total correlation, an individual test item is correlated with overall performance on that test across individuals. If that item is good, then there should be a high correlation, as people who get this item right should score higher on the test. A person total correlation (PTC) is calculated by transposing the item by person matrix used for an item-total correlation, so the responses provided by each individual for each item on a scale are correlated with the sum scores of all other individuals on each item of that scale (Curran & Denison, 2019; Dupuis et al., 2019).

An example of this can be seen in Table 1. Column one indicates the item that an individual is responding to, the second column indicates responses to these items by an individual, and the final column contains the sum scores of all other individuals. Columns two and three are correlated for every person in the dataset and the resulting correlation coefficient is their person total correlation.

The two tables contain example numbers for different participants in a dataset. Notice that the individual responses to each item for each person and the sum scores change. This change reflects the fact that the sum score includes every person in the dataset minus the individual that is currently being examined to avoid artificial inflation of the correlation.

Table 1. Person Total Correlation for Person One and Person Two

Item Number	Individual One	Sum Score
1	5	56
2	4	62
3	5	76
4	5	84
5	3	40
6	4	51

Item Number	Individual One	Sum Score
1	2	59
2	2	60
3	1	72
4	2	81
5	1	38
6	1	48

In the case of testing data, individuals who had a negative person total correlation would represent those who answered high difficulty items correctly while incorrectly answering low difficulty items, suggesting guessing or aberrant responding (Donlon & Fischer, 1968). In the case of polytomous data, negative person-total scores would represent a participant who strongly agreed with some items measuring a psychological characteristic while strongly disagreeing with other items that measure that same characteristic (Curran, 2016; Curran & Denison, 2019; Dupuis et al., 2019). E.g., a participant strongly agrees with the item "Tends to feel depressed, blue", but disagrees with the item "Often feels sad" or strongly agreeing with both "Is emotionally stable, not easily upset" and "Is temperamental, gets emotional easily". It is somewhat unclear how well this method translates from a unipolar construct, like difficulty, to a bipolar construct, like extraversion, however simulation results indicate that this method is effective (Curran & Denison, 2019).

Other Metrics

Several other statistical metrics exist that are not discussed in the current paper (Curran, 2016). Resampled internal reliability and odd-even correlations calculate the reliability of a participant's responses. Odd-even correlation is similar to resampled internal reliability but correlates the odd and even items of a scale, whereas resampled internal reliability randomly resamples split-halves of the scale and averages them. However, because reliability is calculated within participants this coefficient can be zero or even negative if participants have low variability in their responses (which would be expected if participants are responding to a unidimensional scale). The traditional approach, while not clearly discussed, is to calculate split halves across scales in a study, then combine these split halves into a single dataset before calculating a correlation coefficient. This generally resolves the issue of low response variability because individuals are expected to vary across the constructs being measured. However, it is not clear whether this assumption generalized to the current study as only two constructs, positive and negative affect, were measured daily. Thus, this metric was not used.

Mahalanobis distance is a multivariate outlier analysis that examines whether the response set produced by an individual is an outlier (not their trait score). While this method is promising, it has received less study than other metrics and therefore has less defined cut criteria. This method also carries normality assumptions and it is unclear how to extend this method to repeated measures data, such as ESM (Curran, 2016).

Chapter 2: Method

Participants and Procedure

Data used in this study are part of the mPerf project[1] and consistent of responses from 428 participants. Four participants were excluded because they provided data for fewer than 7 ESM periods, leaving 424 participants. For the ESM portion, respondents provided data for a minimum number of 7 days (first quartile 61 days) and a maximum number of 117 days (third quartile 71 days). The average number of days someone provided data for was 64.56 and the median number of days was 66. Most participants began the study in early 2017, but recruitment for some participants continued until 2018. During the ESM portion of the study 27,308 complete response episodes were recorded[2]. Participants were paid for their participation in the study and provided with a cellphone to use when answering surveys for the study.

Most participants were male (58.9%), 39.7% were female, and 2 participants (0.5%) did not respond to this question. Participants were 31.77 years old on average (median age 30) and ranged from 19 to 65 years old. Most participants had a bachelor's degree (45.8%) or master's degree (25.9%), with a smaller portion having a doctoral degree (9.8%), some undergraduate education (8.9%), some graduate education (6.8%), a high school degree (1.2%), or some high school (0.2%).

[1] Data are part of mPerf (http://mperf.md2k.org/), a large interdisciplinary project that uses sensors and software to predict employees' individual differences and work behaviors. Additional details are also provided in (Wiernik et al., 2020).
[2] A small number of response episodes (164) were excluded for not being completed, meaning they did not submit the survey after opening.

Measures

Baseline Measures

Before beginning the study, individuals responded to the BFI-2 (Soto & John, 2017), PANAS-X (Watson & Clark, 1994), the 20 trait items from the state-trait anxiety inventory (Spielberger, 1983), and several other individual difference and job-related measures that will not be used in the current study, e.g., cognitive ability and job performance. This assessment was conducted in person and involved a research manager explaining various aspects of the study and survey before participants responded to these measures using a computer. Among the substantive measures at baseline, there were only 8 missing values. As this number was small, case-wise deletion was used if a participant had missing values on any scales used in a particular analysis. E.g., a participant missing responses for the agreeableness measure was not included in statistical analyses using that variable.

Daily Measures

Every day, individuals received a prompt on their cellphone in the morning, midday, or evening and had twenty minutes to respond to this prompt. The daily measure always included 10-items from the PANAS (Kercher, 1992), a single-item stress measure, a single-item anxiety measure, a single item alcohol-use measure, a single-item tobacco use measure, and a sleep measure. An additional set of items was also randomly chosen from the following each day: self-reported job performance (technical, OCB, and CWB), the BFI-10 (Rammstedt & John, 2007), or two subscales from the BFI-2 (Soto & John, 2017).

Similar to baseline, few missing values were present, with a total of 84 missing values across all responses to the PANAS measure. Given that daily measures were only used to flag participants as careless missing daily data presents less of an issue than missing data at baseline. However, it is difficult to produce scores on some careless response metrics if an individual has missing data. For example, if a participant is missing one value it is hard to produce a score on ISD, as their score may

be artificially inflated or deflated simply because they are missing a value. Standard practice is to score these instances as 'not careless' on these metrics in the hope that other metrics not sensitive to missing values, such as response time, will accurately flag individuals as careless if they indeed appear to be (Curran, 2016)[3].

Screening on Careless Response Metrics

For the purposes of the current study 12 items will be used to screen for careless respondents at each ESM period, the single item stress measure, single item anxiety measure, and 10-item PANAS. This decision was made for several reasons: 1) the PANAS and two single item measures were the only measures that participants responded to every day, 2) the psychometric properties of the PANAS allow for the calculation of the metrics described above and have theoretical relationships with stress and anxiety, and 3) a variant of the PANAS was also used by Jaso et al. in their study.

For the ESM portion, careless response metrics will be calculated for each individual response episode. E.g., "person 1" will have a score on each of the careless response metrics for day 1, day 2, day 3, etc. Because of this, every individual will have multiple scores on each metric corresponding to each day they responded, which can be used to investigate variability in careless responding over their entire ESM period. A rolling window of cutoff levels for each careless metric was used during the ESM period and was chosen to reflect conservative to liberal cut points. The effect that excluding participants based on these cut points has on pairs of highly correlated items was then examined to determine at what cut point these correlations stabilize. Additionally, the positive affect subscale was always assessed before the negative affect scale and this scale was not

[3] Methods such as imputation are generally not considered before careless response screening because the imputation process could directly affect scores on a careless response metric. Consider the case of longstring, a imputed value in the middle of the scale could vastly inflate or deflate a longstring score depending on whether it matched with the existing response string or not.

randomized. The relationship between the last item in the positive and first item in the negative scale will also be assessed. These items should theoretically have no correlation (Kercher, 1992), but careless respondents may fail to notice when these scales shift and this will artificially inflate the correlation between these items.

Cut Scores

Response Time

For response time, cut scores of [≤ 1s, 1.5s, 2s, 2.5s] were used. It is impossible to determine a universal cutoff for response time as a more complex item will take longer to respond to than a simple item (Curran, 2016). Huang et al., (2012) suggest 2 seconds as a cut score that is applicable to many psychological measures and given the simple format of items on this survey this cut score seems appropriate. The above cut scores were chosen to allow for the examination of a window around this 2s cut score, with the cuts below 2s representing a more conservative approach to careless response detection. Given how short the PANAS items are, even the 2s cut score could potentially be a liberal estimate. Since time per-item is not recorded individually, page submit times were summed across the anxiety, stress, and PANAS items and divided by 12 to approximate the average response time per-item.

Longstring

For longstring cut scores during the ESM portion [≥ 8, 7, 6, 5] will be used. Similar to response time, longstring is dependent on the content of the items in the scale, as a scale of 12 highly similar items should have less response variance than a scale of 12 items that each measure a different construct. It is important to note that the anxiety and stress items were asked first, then the positive PANAS items, then negative PANAS items. Because of this we might expect the maximum longstrong for a non-careless participant to be five. Given that each PANAS subscale contained five items, this would represent a participant who responded consistently to items within a subscale but

changed their responses after switching subscales. The stress and anxiety items cannot add to the negative affect longstring because they were separated by the positive PANAS items.

Inter-item standard deviation/Intra-individual response variability

There are no established cut scores for flagging consistent responses with IRV. In their original paper, Dun et al. presented scores using standard deviations from the mean response deviations and arbitrarily cut the worst 10% of respondents. Because of the lack of established cut scores, a window was created around a theoretical participant who was largely invariant but could theoretically be responding conscientiously. Consider the response string: [2,2,1,1,1,1,1,2,2,2,2,2]. This reflects a participant who responded identically to all positive affect items and identically to all negative affect items, plus the anxiety and stress items. This participant shows some variability across constructs, even if that variability is minor. The standard deviation for this response string is 0.51, which could be viewed as the minimum standard deviation a participant can have while still showing differing response patterns between constructs. As such, cut scores constructed around this window [≤ 0.65SD, 0.55SD, 0.45SD, 0.35SD] was used.

There are also no established cut scores for flagging inconsistent responses using ISD. Marjanovic et al. recommend simulating random data that reflects purely random responses to the scale and cutting responses that have higher standard deviations than this simulated data. Based on a simulation of ten million samples randomly drawn from five response options, this standard deviation is 1.41, with a maximum possible standard deviation of 2.19. Thus, cuts between this range were used [≥ 1.8SD, 1.6SD, 1.4SD, 1.2SD] for the ESM portion.

Person total correlation

While strict cut-scores for person total correlation (PTC) have not been recommended, researchers agree that negative scores indicate aberrant responses (Curran, 2016; Donlon & Fischer, 1968; Dupuis et al., 2019). The current study used cut-scores of [≤ -.20, -.10, 0, .10]. In this case, .10

acts as liberal cut-point as it is slightly above the negative threshold traditionally recommended and the 0 and -.10 reflect cut-scores that align with standard recommendations for person total correlation. Last, the -.20 cut reflects a more conservative cut-score that might be used if researchers are worried about inconsistencies in responses that are due to the short length of the PANAS scale used.

Validation Check with Psychometric Synonyms

As discussed above, the PANAS measure used in this study contained no psychometric antonyms, but did contain three psychometric synonyms, anxious-nervous, scared-afraid, and excited-enthusiastic. Due to the small number of these pairs, they were not used to directly detect careless responding; instead, the correlation between these items served as a criterion for evaluating whether the respondents removed appeared to be careless. Specifically, the relationship between item pairs for people flagged as careless at each cut point was compared to the relationship for people not flagged. The relationship should be weaker for inconsistent careless respondents and stronger for consistent careless respondents (Jaso et al., 2021). As an additional screen for consistent respondents, the correlation between "determined" and "distressed" was examined as this represents the point in the PANAS scale where items switch from positive to negative affect. The correlation between these items should theoretically be close to zero, e.g., in the original validation study this correlation was .06, so a correlation between them in the careless sample may represent "drift" as they switch from responding to one construct to another (Kercher, 1992).

This pairs will be used to select "optimal" cut scores on each metric that will be used to flag respondents as either carless or not careless in an overall flag. The word optimal is not meant to insinuate that these cut scores are truly optimal, but that they represent the integration of theoretical cut scores with empirical data to produce a best guess at what an optimal cut score would be. This

also simplifies the analysis process for later questions, as fitting and reporting all regression models for every combination of cut scores would be expensive in terms of time and space.

Predictors of Careless Responding

To examine potential predictors of careless responding generalized additive mixed models were fit with a dichotomous outcome of whether a participant was careless or not for a given response episode using the optimal cut scores determined above. Mixed models with random slopes and intercepts were chosen to account for dependencies in careless responding within individuals and to allow for the examination of trajectories of carelessness within individuals over time. Theoretically, each individual has a unique propensity to respond carelessly at the first timepoint and may have different trajectories in their careless behavior, so random slopes and intercepts are both warranted. Generalized additive models (GAMs) were chosen instead of linear models because they allow for a great deal of flexibility in the shape of a regression and are ideal for modeling processes that unfold over time in non-linear patterns. It is likely that the propensity to respond carelessly does not follow a simple linear trend over time and instead has a non-linear increase and decreases for different response episodes. As an example, consider a participant who is generally conscientious in their responses but has an extremely busy week at work and responds carelessly during that time. A GAM allows for their propensity to respond carelessly during that time to increase before decreasing again.

An additional benefit to GAMs is that the flexibility of the line fit by a generalized additive model is controlled using a penalty parameter. This penalty can shrink the "wiggliness" of the GAM to zero, producing a linear line, if responses approximate a linear pattern. Thus, generalized additive models are especially useful when the form of the regression line is not clear a-priori, which would make fitting more prescriptive models, such as a polynomial or piecewise regression, difficult (Baayen & Linke, 2020; Wood, 2017).

To begin, a baseline GAM model was fit with random slopes and intercepts for participants across time and a fixed effect of time. Using standard linear notation, the equation for this model is displayed in (1). To modify this model from a generalized linear mixed model to generalized additive mixed model, the beta coefficients are replaced by a function of x that determines the degree of smoothing by adding together basis functions of x with a penalty term to avoid overfitting. For the sake of brevity, all models are presented using R code in Appendix A. For a detailed explanation of generalized additive modeling using R see Wood (2017).

$$\text{Careless}_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}(p_{ij}) = \beta_{0j} + \beta_{1j}\text{Time}_{ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\begin{bmatrix} \mathbf{u_{0j}} \\ \mathbf{u_{1j}} \end{bmatrix} \sim \mathbf{MVN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\tau}_{\mathbf{00}}^{\mathbf{2}} & \boldsymbol{\rho\tau_{00}\tau_{11}} \\ \boldsymbol{\rho\tau_{00}\tau_{11}} & \boldsymbol{\tau}_{\mathbf{11}}^{\mathbf{2}} \end{bmatrix} \right) \quad (1)$$

To fit the "random" portion of the model, factor smooths were used for each participant in line with suggestions from Baayen & Linke (2020). This allows for the examination of trajectories in individual carelessness across time. In addition to examining individual trajectories, the overall relationship between time and careless response rates was modeled. Baayen & Linke (2020) suggest that a fixed effect term can be included in addition to the random effect covariate in the factor smooth, in this case time, if there is theoretical reason to believe that people will follow a general trend over time in addition to specific, individual trends. In this case, as stated by research question four, we believe careless response rates will generally increase with time.

Finally, the type of smoother for these effects must be determined. In the model containing only time, thin plate splines were used as the smoothing method. However, thin plate splines cease to

become effective for smoothing when interactions between variables that share different scales are introduced to the model. As time and personality scores do not share a common scale, tensor product smooths were used for models incorporating both variables (Wood, 2017).

Chapter 3: Results

All analyses were conducted in R (R Core Team, 2019). To run careless response detection metrics, modified functions from Curran (2018) were used. Analyses were conducted using the lme4 and glmmTMB packages for mixed effects modeling and the mgcv package for generalized additive models (Bates et al., 2015; Brooks et al., 2017; Wood, 2011; Wood et al., 2016). The tidyverse family of packages was used for data manipulation and plots were produced using ggplot2 (Wickham et al., 2019). Finally, the see, modelbased, parameters, and performance packages from the easystats family were used to produce parameter estimates, conduct model assessment, and to assist with visualization (Lüdecke, Ben-Shachar, et al., 2021; Lüdecke et al., 2020; Lüdecke, Patil, et al., 2021).

Examination of Cut Scores

The distribution of each careless response metric is visible in Figure 1. The response time graph was modified to exclude response times greater than 20 seconds per-item as cutting excessively long response times tends to already be standard practice. From this graph, it is apparent that the mode longstring is 5, potentially demonstrating that responding identical within constructs is a common phenomenon. There is also a spike at longstring scores of 12, which denotes response episodes that were identical responses were selected for every question. This corresponds to the large number of response episodes with 0 standard deviation between constructs, whereas any longstring of 5 within a construct also potentially represents a standard deviation of 0 within constructs. Also, note the limited range of the standard deviation within constructs. Few response episodes had a standard deviation greater than 1. Finally, the PTC plot show that most episodes showed consistency, however there is a long tail of episodes with negative PTC. Given that this tail extends well beyond

the -0.20 cut point used in this study, researchers could potentially employ even more conservative cut scores should they wish to.

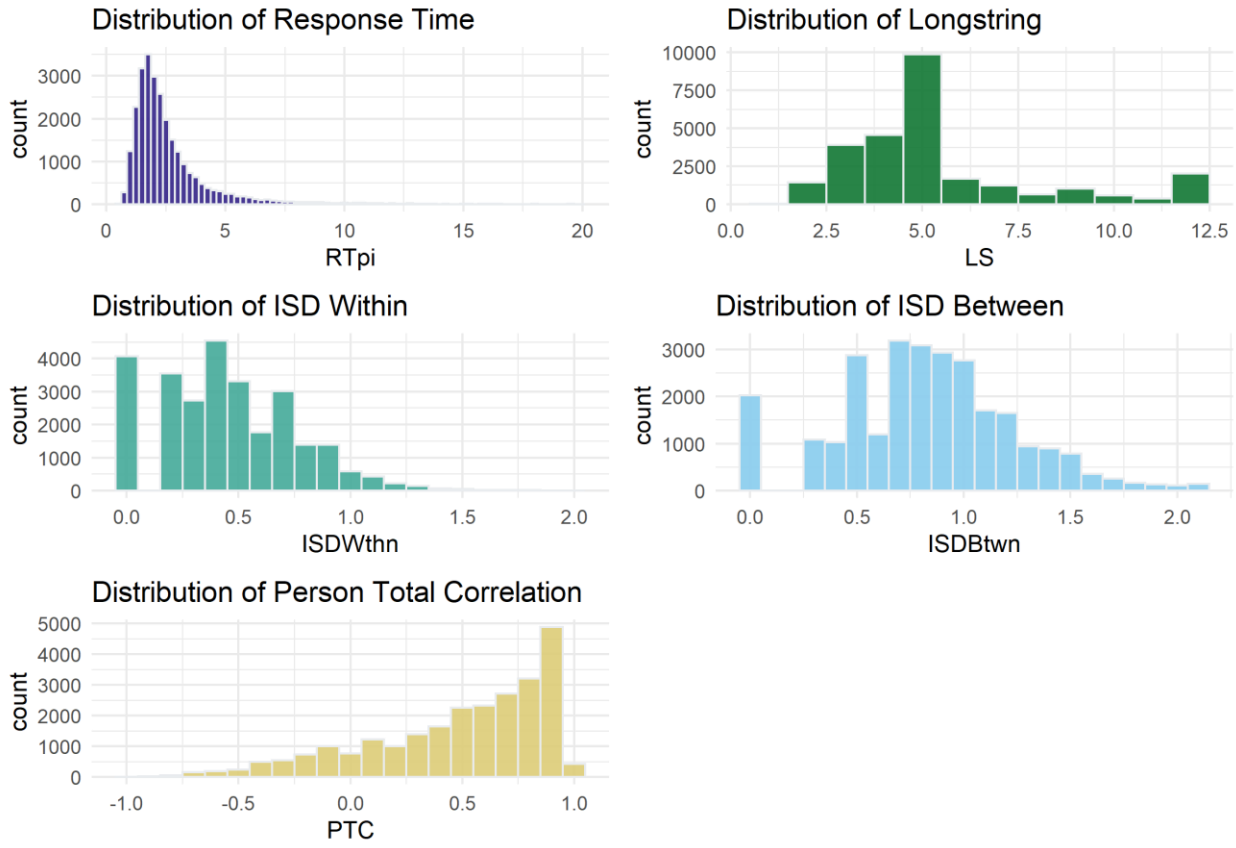

Figure 1. Distribution of Careless Response Metrics

Correlations among the psychometric synonyms and the distressed-determined item pair were examined after response episodes at each cut score of each careless response metric were flagged independently. The correlation among items for each cut point are compared to the participants flagged as not careless in Figures 2-6. It is important to note that each of these groups are mutually exclusive. That is, respondents visualized as having a response time of ≤ 1s are independent of respondents flagged as ≤ 1.5s. In the visualization, ≤ 1.5s respondents have response times 1s < and ≤ 1.5s. This categorization was used to more accurately examine response trends at each cut point. If a cumulative flagging was used instead, it could result in a case where moving from one cut

point to another appeared to have no effect on item correlations simply because cases at previous cut points overpowered this new data.

Figure 2 displays the relationship between item pairs for each corresponding response time cut value. As a reminder, nervous-anxious, scared-afraid, and enthusiastic-excited are expected to have positive relationships while distressed-determined are expected to have no relationship. This is the case for each cut point except for the response time ≤ 1s cut point. In this case there is a clear positive relationship between the distressed-determined item pair and the regression line for this group also shows small deviations from the overall regression line for other items. A cut score of under one second also algins with Jaso et al., so was chosen as the response time flag.

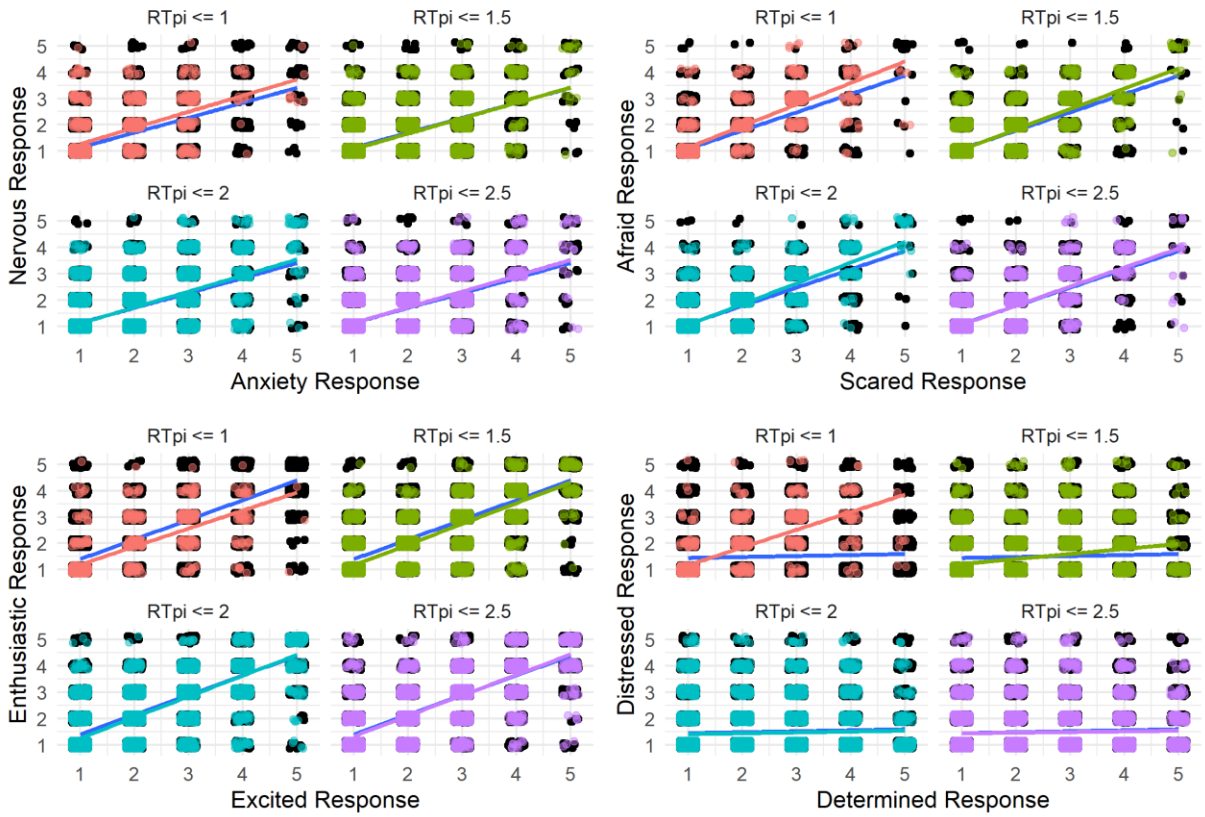

Figure 2. Relationship Among Criterion Items at Response Time Cut Points

Note. Dark blue line and black points represents all data not flagged and colored lines and points represent flagged data. Due to scaling, single points on the graph represent dozens of responses for that particular response option combination

Next, correlations among these items were examined at each longstring cut point and are displayed in Figure 3. Again, the largest differences in item correlations can be seen for the distress-determined item pair. In the non-flagged sample this relationship is almost zero, however for longstring scores of 6 or above there is a positive association between these two items. This relationship could be explained if these participants adopted a strategy of responding identically within PANAS subscales, but occasionally failed to modify their response when switching constructs. Additional analysis identified 4062 (14.87%) response episodes that were flagged as having a longstring of 5 on both the positive and negative PANAS items. The number of response episodes where a participant responded identically to all PANAS items was 2871 (10.51%), leaving 1,191 (4.36%) episodes where respondents had identical responses within subscales but modified their responses between subscales.

Thus, while this strategy was not the most popular, it does appear that a substantial number of response episodes followed this pattern of identical responses within PANAS subscales. A cut point of Longstring ≥ 6 was chosen at it represents the clearest difference between relationships on the criterion items, however a finer grained analysis might also include the 4.36% episodes discussed above. This cut score is also similar to that of Jaso et al., who used a cut score of 60% of responses at the mode.[4]

Correlations among items for each ISD between constructs cut window are displayed in Figure 4. Similar to the above metrics, the distressed-determined pair shows the largest difference in item correlations and is consistently positive at all cut values. This cut score is above the simulated minimum cut score of 0.51, but still represents little variability between responses. Examining the distribution of ISD between in Figure 1 there is a visible spike in the histogram at this 0.50 mark.

[4] This study employed response sliders instead of Likert type responses, so cut scores differed in their metric.

However, values around 0.60 also show a departure from the rest of the distribution. Because of this, and the evidence from the distressed-determined item pair, the ≤ 0.65 cut point was chosen.

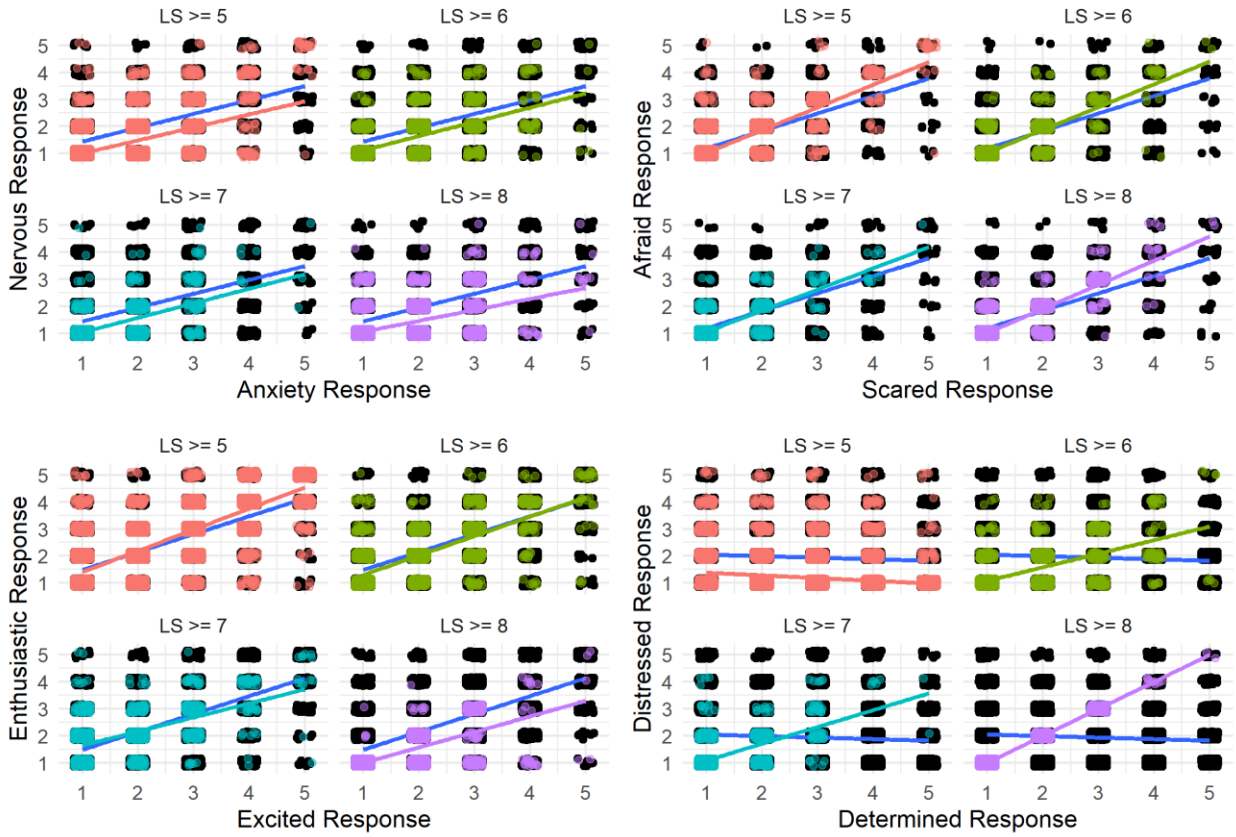

Figure 3. Relationship Among Criterion Items at Longstring Cut Points

Note. Dark blue line and black points represents all data not flagged and colored lines and points represent flagged data. Due to scaling, single points on the graph represent dozens of responses for that particular response option combination

In contrast to the above metrics, ISD within constructs represents the first metric designed to detect overly inconsistent responding. However, the relationship between this metric and the validation items is less clear cut than the above examples. The window of [1.2, 1.4, 1.6, 1.8] was examined, however no clear differences in the relationship between items is apparent, except at the 1.8 cut, which appears to be driven partly by the lack of available datapoints (only 5 episodes were flagged at this cut score). This lack of data is also present at the 1.6 cut score, with only 36 episodes

being flagged. Similarly, a cut of 1.7 also only produced 19 flagged episodes. Because of this lack of data an optimal value for selecting a cut score is still unclear. Examining Figure 1, very few values are present on this metric before the 1SD point. However, this represents a point well below the 1.3SD produced by simulating random data. Because of the lack of extreme values on this metric, a conservative value of 1.6 was selected as this would correspond to a participant selecting a unique value for every response or using opposite extreme ends of the scale[5].

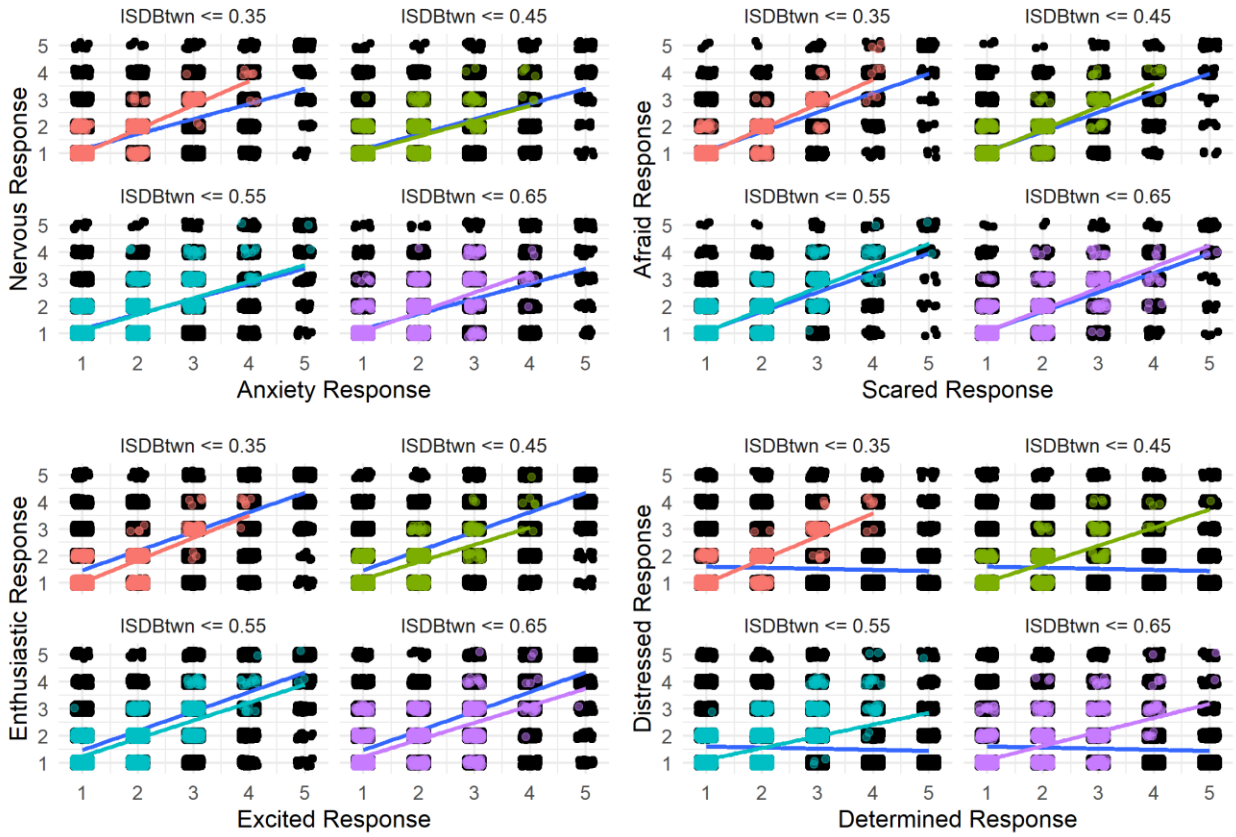

Figure 4. Relationship Among Criterion Items at ISD Between Cut Points

Note. Dark blue line and black points represents all data not flagged and colored lines and points represent flagged data. Due to scaling, single points on the graph represent dozens of responses for that particular response option combination

[5] The standard deviation of the set [1,2,3,4,5] is 1.58, however this cut score excluded the same number of participants as the 1.6 cut.

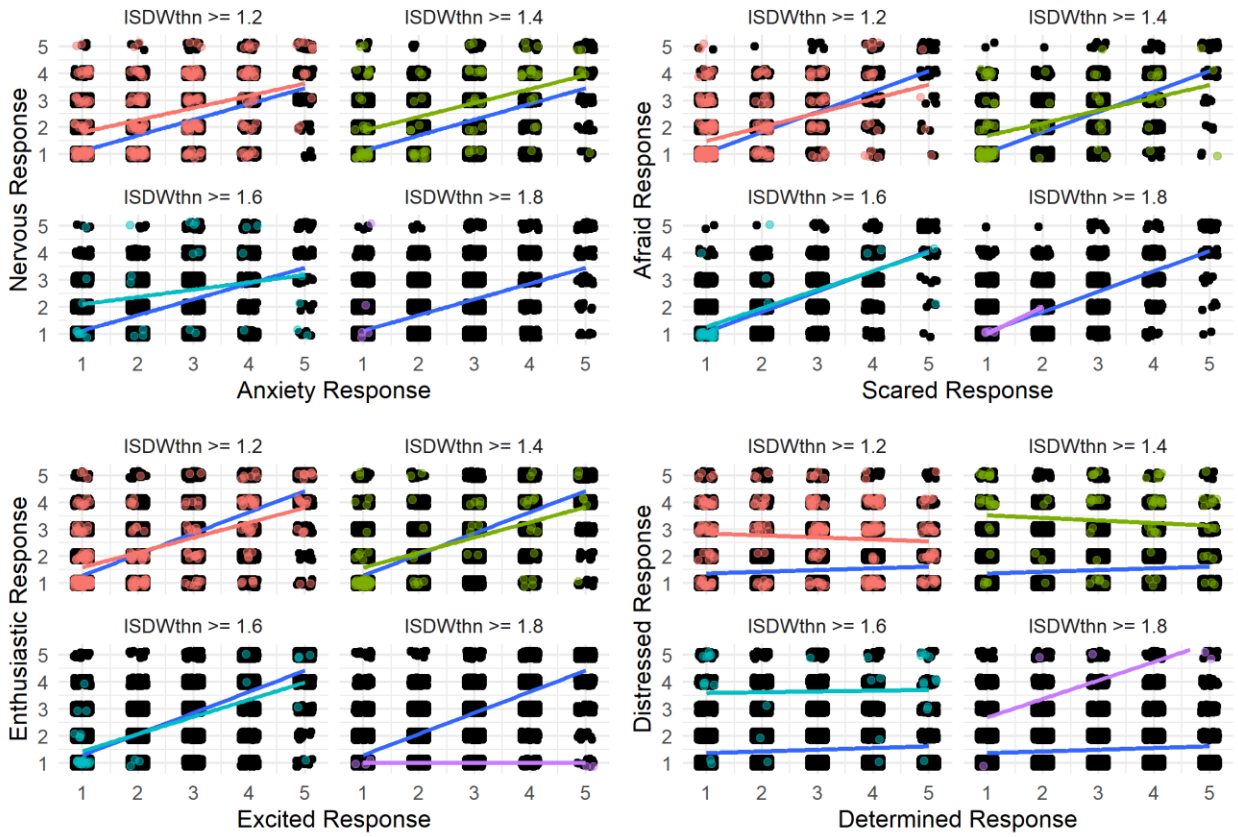

Figure 5. Relationship Among Criterion Items at ISD Within Cut Points

Note. Dark blue line and black points represents all data not flagged and colored lines and points represent flagged data. Due to scaling, single points on the graph represent dozens of responses for that particular response option combination

Finally, the relationships among items at each person total correlation (PTC) cut are presented in Figure 6. PTC has previously been discussed as a method for detecting overly inconsistent respondents (Curran, 2016), however Figure 6 indicates that in this sample it is also detecting overly consistent respondents, given that item relationships are stronger among those flagged by person total correlation. As discussed above, this is because PTC acts as a type of outlier analysis, in that individuals who respond to items in a way that is inconsistent with how other people respond are flagged. In this case, it appears that some of that inconsistency arises from participants who are overly consistent. In the below analysis, the cut point of ≤ 0.10 shows divergence from the reference line for the distressed-determined item pair, however, this point shows less divergence

from other item pairs. At the cut point of ≤ 0 there is clear divergence from the reference line for the overall sample, thus this cut point was chosen as it also aligns with suggestions from prior literature (Curran, 2016).

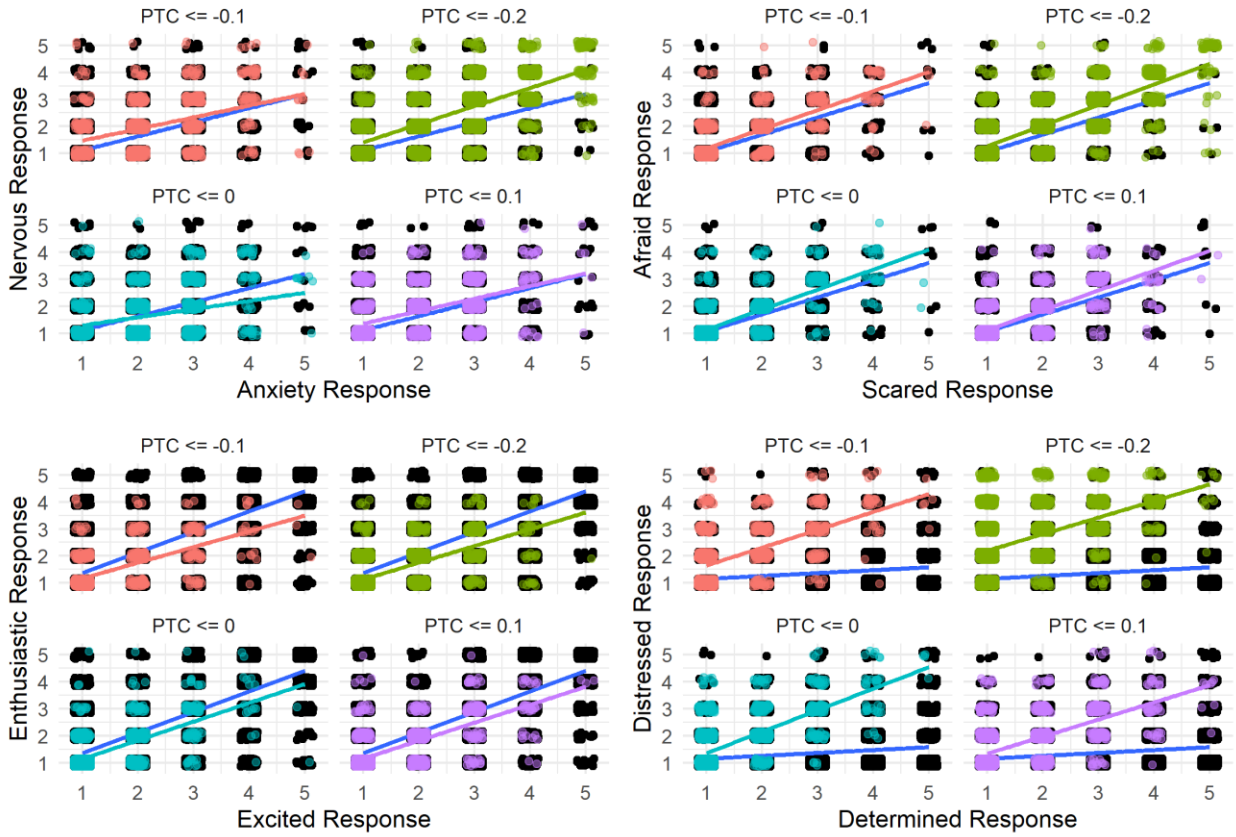

Figure 6. Relationship Among Criterion Items at Person Total Correlation Cut Points

Note. Dark blue line and black points represents all data not flagged and colored lines and points represent flagged data. Due to scaling, single points on the graph represent dozens of responses for that particular response option combination

It is also important to note that because PTC detects outlying response patterns that it is possible it flagged participants who had rare but valid response patterns. In particular, note that PTC had a tendency to flag participants who reported high levels of distress, likely because this response option was rarely selected. This hypothesis was confirmed by visually examining response patterns for episodes where 5 was selected. Often, these participants were flagged as careless and truly appeared to be. That is, they showed inconsistency in how they responded to other similar items on

the negative affect scale or showed longstring behavior. However, some participants exhibited high levels of distress and appeared to respond in a manner consistent with that standing but were still flagged as careless.

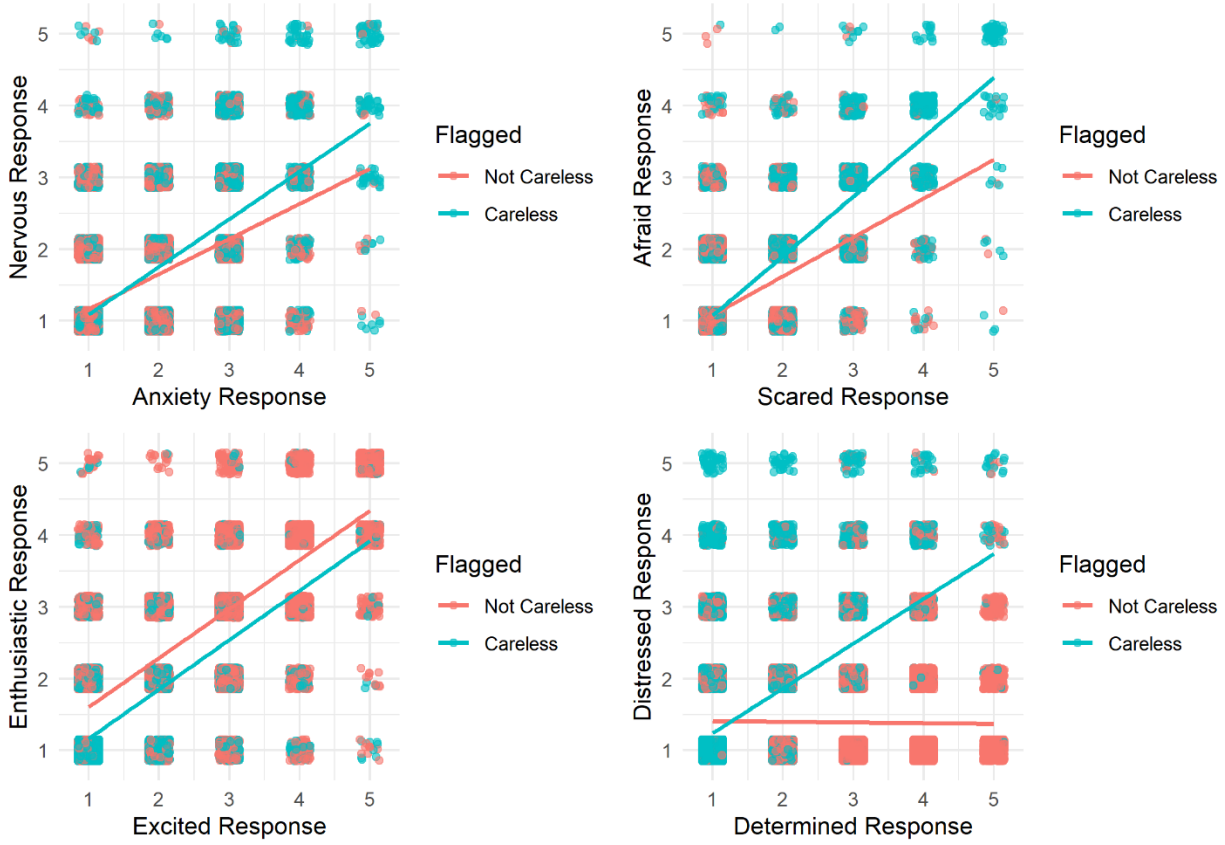

Figure 7. Relationship Among Criterion Items After Flagging

This is obviously undesirable behavior on the part of this metric, but it highlights a point that has so far not been discussed; each of these metrics have false positive and false negative rates. In this case, the false positive rate for PTC appears to be related to selecting rare response options, such as 5's on distressed, scared, afraid, etc. This behavior may not occur in samples that do not measure constructs that occur infrequently, such as fear, but should be kept in mind as a possible limitation of the utility of PTC. Additional methods could be employed to account for this behavior, such as manually examining cases where rare response options were selected or excluding these cases from PTC analysis in the hope that other metrics function, but these intricacies were beyond the scope of

the current paper. However, in order to consider the effect this behavior could have on analyses, some results are reported excluding PTC.

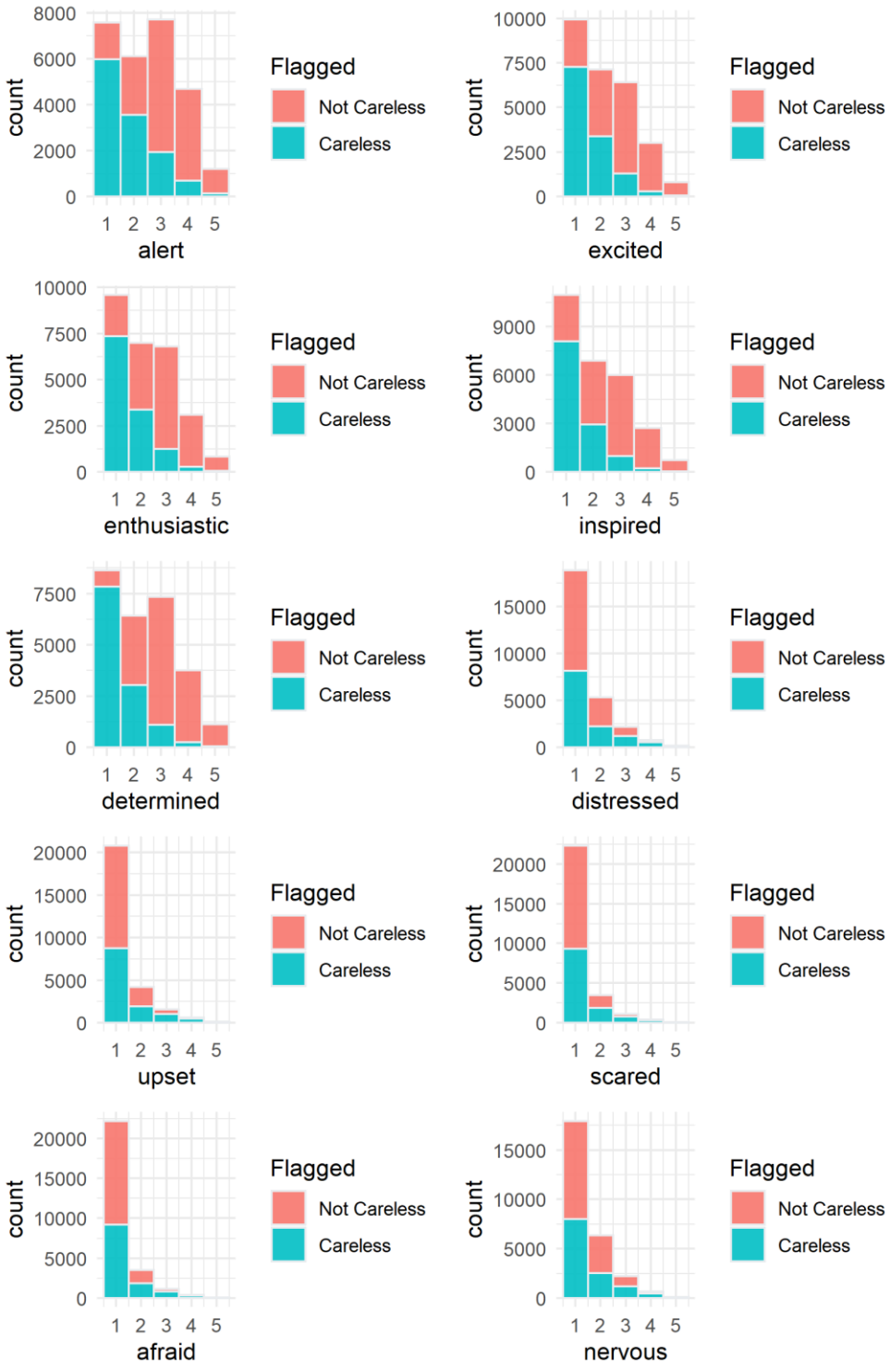

Figure 8. Distribution of PANAS Item Responses After Flagging

The correlations among items for those flagged with the combination of all cut scores [RT ≤ 1s, ISD-B ≤ 0.65, PTC ≤ 0, longstring ≥ 6, ISD-W ≥ 1.6] is displayed in Figure 7. As can be seen, the difference between slopes among the careless and non-careless groups is noticeable for all items and especially drastic for the distressed-determined pair.

In addition to the correlations among item pairs, the distribution of responses to each item is displayed in Figure 8. One pattern in the flags that is visible is that responses of 1 to the positive PANAS items tended to be flagged frequently as careless, whereas this was not the case for negative items. It is not immediately clear why this is the case, but one potential explanation might be that participants responded with longstrings at the start of the item block and switched responses towards the end in an attempt to not appear careless. It could also be the case that many participants were simply low on negative affect, so showed high levels of consistency within these items.

Prevalence of Careless Responding

To answer research question one about the prevalence of careless respondents, the percentage of response episodes flagged as careless was calculated using the most conservative and liberal values on each cut score in addition to the cut scores selected above using criterion item relationships. In the most conservative scenario where all cut scores were at their minimum, this was 25.4% (maximum and 3rd quartile = 100%). Using the optimal cut scores selected above this was 44.98% (if PTC was not included the percent flagged was 38.25%) and using the most liberal cut scores on each metric, except for ISD within, which was restricted to 1.4 due to their being little support for even this cut point, the proportion of episodes flagged was 88.4%[6].

[6] It is important to note that this liberal value is likely an overestimate given that response times of 2.5 are likely still conscientious. Reducing this value to 1.5 flags 77.65% of responses instead and setting a longstring cut of 6 further reduces this to 52.09%.

Second, the percentage of participants who had ≥ 25%, ≥ 50%, ≥ 75%, and ≥ 90% of their response episodes flagged as careless was calculated using the optimal cut scores selected above. The percentage of participants who had 25% or more of their response episodes flagged was 71.0%, for 50% or more episodes it was 40.1%, for 75% or more it was 19.6%, and for 90% or more it was 7.1%. Finally, 1 participant was flagged as careless at every response episode and 11 participants never responded carelessly[7]. This suggests that, while most people are sporadically careless in their responses, there are some frequent offenders who repeatedly provide careless data. It also illustrates that almost every participant was careless several times during the study.

To further investigate whether participants continued to be careless once they were careless the first time, the proportion of response episodes that were flagged as careless after a participants first careless response episode was calculated. On average, 47.97% (median = 44.82%) of response episodes after the first instance of carelessness were also flagged as careless (min = 1.39%, 1st quartile = 24.62%, 3rd quartile = 71.64%, max = 100%, sd = 27.90%). Thus, while there appears to be some consistency in careless responding after the first episode, there is considerable variability, with some participants always responding carelessly after this first point while others only responded carelessly once or twice

The total number of participants flagged as careless 90% or more of the time after their first careless response episode was 38. Most of these participants responded carelessly long before the study ended, mean days careless = 62, median = 66, minimum = 19, maximum = 77. That is to say, those being flagged as repeatedly careless after their first episode were not just careless in their last day or two of the study. Instead, they showed consistent patterns of careless behavior well before the study ended and, in some cases, for nearly all of their response episodes.

[7] If the cut score was relaxed to not include PTC 17 people were never flagged as careless.

Agreement Among Careless Response Indices

Careless response metrics that detect similar response styles are expected to have some overlap in which participants are identified as careless. The number of response episodes flagged by each metric is displayed in Table 2 and shows both unique and overlapping flags. As can be seen, a large number of response episodes were flagged by longstring an ISD. Together these metrics flagged 15.08% of the response episodes as careless. A large number of these are the result of identical response being selected by every participant, which, as discussed above, happened in 10.51% of response episodes. Outside of this agreement, both methods each flagged 6% and 7% of episodes uniquely. In the case of ISD these are response episodes that were highly consistent but varied their responses enough to not meet the 6 longstring threshold. In the case of longstring this represents episodes where 6 or more identical response options were selected, but still had high standard deviations. This likely occurred because participants selected extreme ends of the scale (e.g., 1 for five items and 5 for five items).

Person total correlation flagged a similar number of respondents uniquely, 6.73%, but demonstrated less overlap with other metrics. While around half of the episodes flagged by this metric were flagged by other metrics, this is smaller than the agreement between longstring and ISD between, where well over half of the response episodes flagged were also flagged by another metric. One reason for this unique flagging could be due to PTC detecting inconsistent responses that were not detected by other metrics. However, this number of unique flags could also be inflated due to false positives, as discussed above. Response time shows a great deal of overlap with longstring and ISD, as well as other metrics to a lesser degree. Finally, ISD within flags few episodes overall.

It is interesting to note that longstring and ISD between make up the bulk of the flags for these cut scores. One could attribute this to the more liberal cut scores used for these metrics, however even at the most conservative cut score of 12 identical responses 10.51% of the episodes

were flagged. This speaks to the fact that this sample exhibited a great deal of overly consistent response behavior.

Table 2. Counts and Percentages of Episodes Flagged by Each Careless Response Metric

Response Time	ISD Within	ISD Between	PTC	Longstring	Number	Percent
Not Flagged	Not Flagged	Not Flagged	Not Flagged	Not Flagged	15025	55.02%
Not Flagged	Not Flagged	Not Flagged	Not Flagged	**Flagged**	1685	6.17%
Not Flagged	Not Flagged	Not Flagged	**Flagged**	Not Flagged	1839	6.73%
Not Flagged	Not Flagged	Not Flagged	**Flagged**	**Flagged**	302	1.11%
Not Flagged	Not Flagged	**Flagged**	Not Flagged	Not Flagged	2060	7.54%
Not Flagged	Not Flagged	**Flagged**	Not Flagged	**Flagged**	4117	15.08%
Not Flagged	Not Flagged	**Flagged**	**Flagged**	Not Flagged	517	1.89%
Not Flagged	Not Flagged	**Flagged**	**Flagged**	**Flagged**	930	3.41%
Not Flagged	**Flagged**	Not Flagged	Not Flagged	Not Flagged	23	0.08%
Not Flagged	**Flagged**	Not Flagged	Not Flagged	**Flagged**	2	0.01%
Not Flagged	**Flagged**	Not Flagged	**Flagged**	Not Flagged	11	0.04%
Flagged	Not Flagged	Not Flagged	Not Flagged	Not Flagged	80	0.29%
Flagged	Not Flagged	Not Flagged	Not Flagged	**Flagged**	21	0.08%
Flagged	Not Flagged	Not Flagged	**Flagged**	Not Flagged	78	0.29%
Flagged	Not Flagged	Not Flagged	**Flagged**	**Flagged**	32	0.12%
Flagged	Not Flagged	**Flagged**	Not Flagged	Not Flagged	59	0.22%
Flagged	Not Flagged	**Flagged**	Not Flagged	**Flagged**	403	1.48%
Flagged	Not Flagged	**Flagged**	**Flagged**	Not Flagged	37	0.14%
Flagged	Not Flagged	**Flagged**	**Flagged**	**Flagged**	87	0.32%

One final surprising finding was that ISD within and longstring had two overlapping flags. Upon further inspection, these cases were individuals who selected 5 for both the first and last PANAS questions, but responded with a 1 to every other question. These are the "alert" and "nervous" items. While this appears to show some consistency, in that someone could have high levels of alertness and nervousness, these participants responded with a 3 and 1 respectively to both the anxious and stressed items. It seems quite unlikely that someone would feel extremely alert and nervous while feeling little to no anxiety or stress.

After metrics were run, the correlation among metrics was computed and these correlations are displayed in Table 3. Scores on the metrics were used instead of the flagging variable to avoid potential inflation or deflation of the correlation due to the possibility that cut scores selected on some metrics were better than others. Unsurprisingly longstring and ISD within and between constructs have a strong negative correlation. Increasing longstring by necessity will decrease the standard deviation of responses. It is interesting that none of the metrics correlate highly with response time, potentially suggesting that speed may not have been the primary motivation for responding carelessly. Finally, note that PTC has a negative correlation with longstring and ISD within, but a positive correlation with ISD between. This indicates that PTC is flagging individuals with higher levels of variability to their responses, suggesting that while it flagged some overly consistent respondents it is primarily flagging inconsistent respondents.

Table 3. Correlations Among Careless Response Metric

	Response Time	Longstring	ISD Between	ISD Within	PTC
RT	1.00				
LS	-0.02	1.00			
ISD-B	0.02	-0.36	1.00		
ISD-W	0.03	-0.56	0.31	1.00	
PTC	0.00	-0.14	0.39	-0.20	1.00

Change in Careless Responding Over Time

To answer research question four, a continuous time variable was created within each individual to track days since they began the study, with day 1 representing their first response episode and day n representing their final response episode. This variable was then regressed on the careless response flag created above using REMAL estimation with thin plate spline smoothing on the fixed effect term and factor smoothing on the random effects.

Using factor smoothed caused issues with model fitting. The factor smooth model was allowed to run for three hours before manually stopping the process. Several adjustments were made to the bam function options in an attempt to aid model fitting, including utilizing parallel processing and modifying several other performance options provided by the bam function, however the model continued to have run time issues. Factor smooths were instead replaced by the "re" argument to fit a model containing random smoothed slopes for participants.

The model using random effects for the smoothing argument ran successfully, however, investigation of the model revealed poor fit, with only 73% of the binned residuals being contained in the error band and significant deviations in the tails of the qqplot. Further, the smoothing value for time was low, 2.8. Smooths for the predictors of conscientiousness and neuroticism were similarly low (1.57 and 2.34 respectively). Autocorrelations were examined to determine if including them in the model would improve fit, however, lag 1 correlations were estimated to be marginal (0.03) as were longer lags. A model containing no smooths was compared to the smoothed time model and produced marginal AIC differences (1.79), with weighted AIC favoring the linear model. Binned residuals were also significantly improved by fitting the linear model, with 84% being contained in the error band. Because of this, a GAM model does not appear to be appropriate.

A linear model was fit instead using the glmmTMB package with random slopes and intercepts as described in (1). This model performed better in terms of binned residuals, with 84% of

the residuals being contained in the error bound. The qqplot also showed no deviations from the fit line, unlike in the previous model. However, as can be seen in Figure 9, which was produced using the default cubic spline smooth in ggplot2, the relationship between time and careless responding does not follow a completely linear trajectory. Specifically, there appear to be two noticeable knots around time 15 and time 60/70 that change the slope of the regression line.

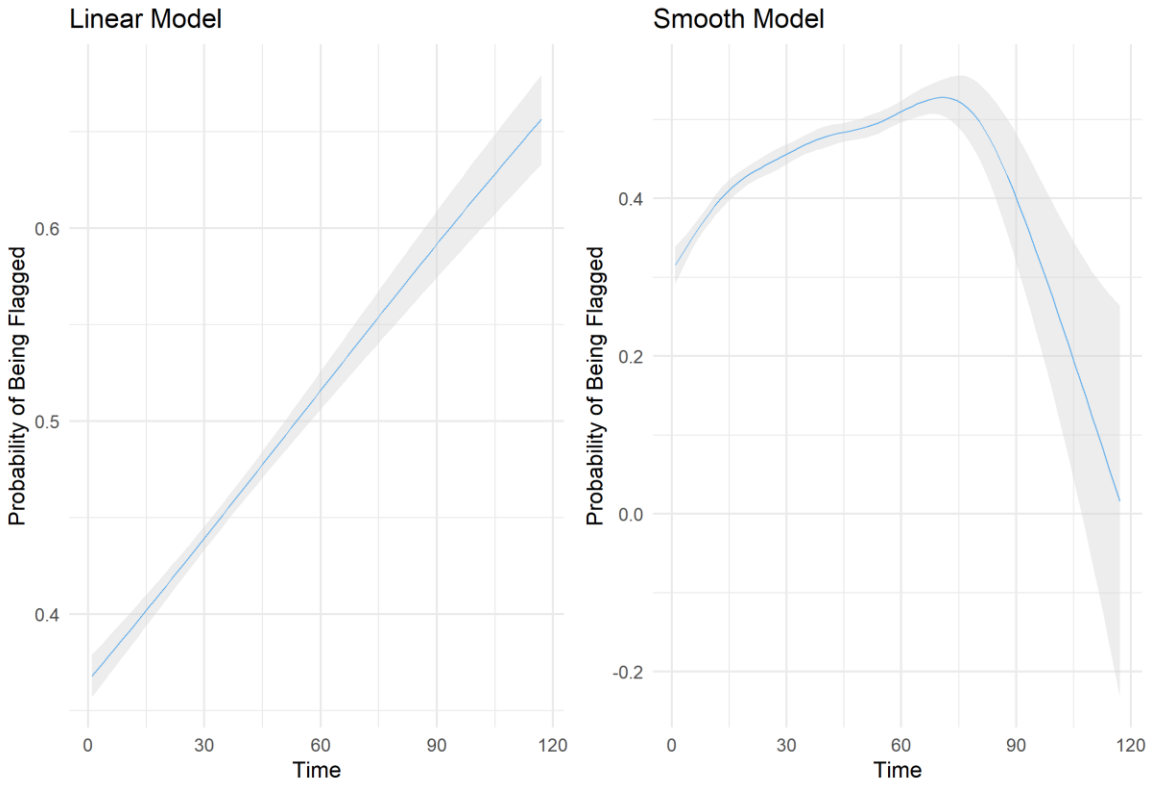

Figure 9. Relationship Among Criterion Items at Person Total Correlation Cut Points

Note. The linear model was fit using the formula lm(y~x) and the smooth model was fit the formula y ~ s(x, bs = "cs").

To account for the major knot, a linear piecewise and cubic spline model[8] were fit with a knot at time 60 to account for the downward trajectory at this point[9]. Compared to the linear model, the

[8] Cubic splines were fit by wrapping the fixed and random effects in separate ns() functions from the splines package.
[9] The reader will note this downward trajectories also corresponds to an increase in the error band. The authors attempted to include a dispersion term in the model to account for this, but the dispersion term was ignored by glmmTMB. It was unclear why this occurred, and time constraints prevented the authors from finding a solution before this paper was complete.

two models with knots at time 60 significantly reduced information loss, with the cubic spline model performing best. Results of this comparison are visible in Table 4. Additional models were run that moved the knot around the time 60 point (e.g., one model set the knot at time 70), however none of these models demonstrated better fit than the time 60 model.

Table 4. Model Comparisons for Time Regression

Model	AIC	AIC Weights	R^2 Conditional
Linear	27856.773	< 0.001	0.54
Linear Spline	27828.717	< 0.001	0.54
Cubic Spline	27761.493	< 0.001	0.54
Cubic Spline Knot at 15	27733.816	0.35	0.54
Cubic Spline Knot at 10	27732.568	0.65	0.54

Note. All models besides the "Linear" model included a knot at time 60.

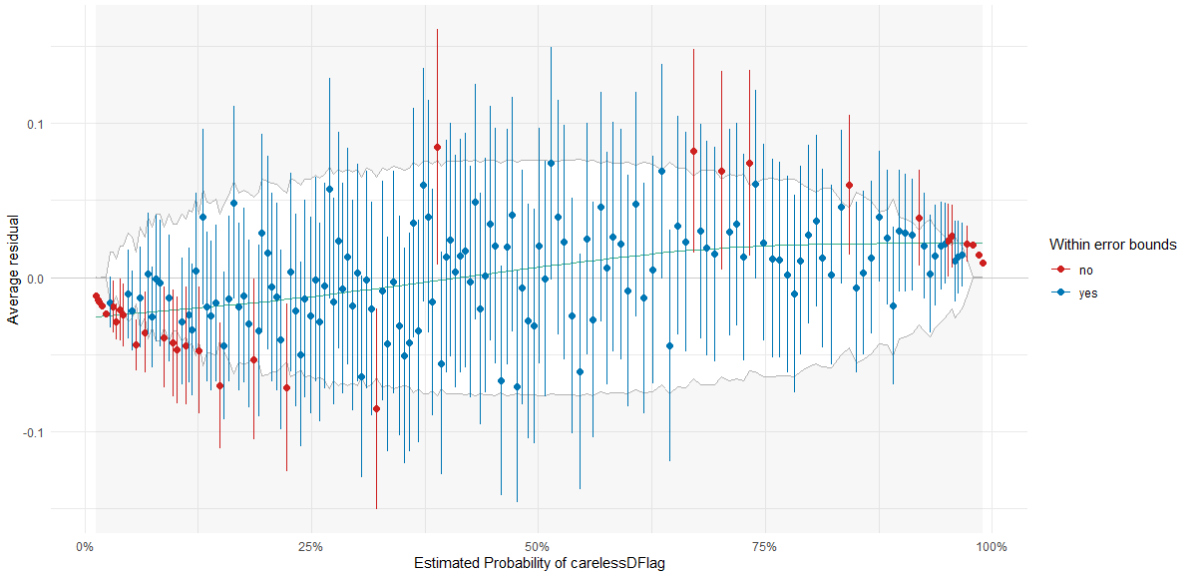

Figure 10. Binned Residuals for Cubic Spline Model with Two Knots

Next, it was tested whether including a second knot at time 10 or 15 improved model fit. Results from this comparison are also shown in Table 4 and indicate that the cubic spline model with a knot at both time 60 and time 10 reduced information loss the most and consequently had the

highest AIC weight. While the information loss reduction was minimal between the 15 and 10 model, there was no other clear reason to prefer one over the other.

Binned residuals for the final cubic spline model with a knot at time 10 and 60 exhibited fair coverage (81% inside error band), with most points departing the error bands at the ends of the distribution (see Figure 10). The qqplot also shows good fit along the predicted line and is presented in Figure 11.

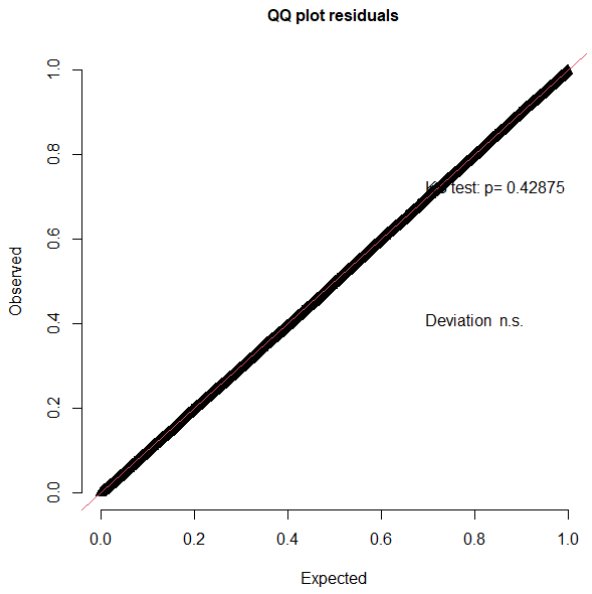

Figure 11. qqPlot for Cubic Spline Model with Two Knots

Results for this model are displayed in Table 5 and visualized in Figure 12. Model parameters are difficult to interpret given the complexity of the model, but Figure 12 provides an overview of the observed relationship. This model shows support for research question 4, that the time in the study has a positive association with the propensity to respond carelessly. There is an initial steep slope for the probability of responding carelessly before flatting off. This indicates that the largest increases in careless responding happen during the first few weeks of the ESM period, with the probability of responding carelessly increasing more gradually after this. Around the day-70 mark the confidence interval becomes increasingly wide due to the sparse nature of the data this late in the study.

However, confidence intervals before this point are quite narrow, demonstrating that even the lowest plausible slopes still correspond to a large increase in the probability of responding carelessly as time increases.

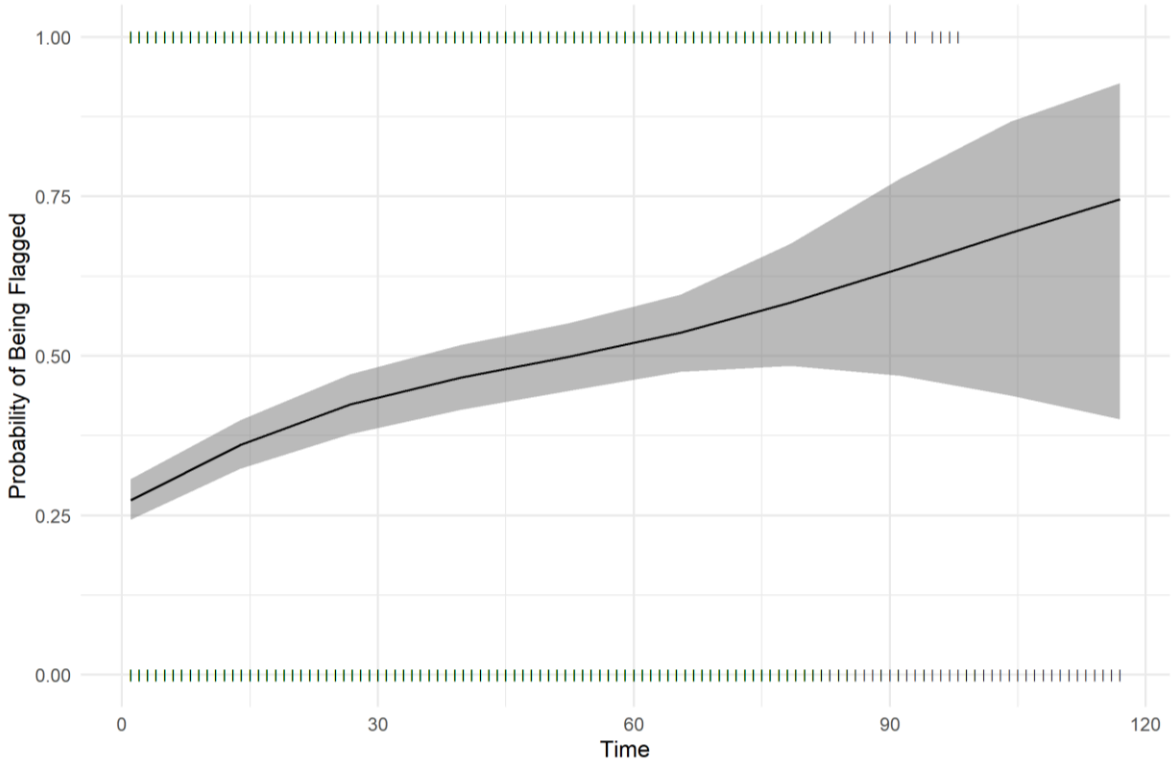

Figure 12. Careless Responding Over Time

Examining the random effects results, it is interesting to note that the intercept and first slope have a negative correlation. This indicates that individuals who had a low probability of responding carelessly at their first episode had a steeper first slope than individuals who had a higher probability of responding initially. Results from the above analysis showing that nearly all participants were flagged as careless for several response episodes (71% were careless for 25% or more of their response episodes and only 11 participants were never flagged as careless) could explain this initial steep slope. That is, because nearly all participants were flagged as careless at multiple episodes even

people who were initially not careless still had several careless episodes later. Correlations between the remaining slopes and intercept are both positive.

Table 5. Results for Time Regression

Parameter	*b*	*b* 95% CI [LL, UU]	SE	p-value
(Intercept)	-0.97	[-1.13, -0.82]	0.08	0.000
Time Slope 1	0.72	[0.27, 1.16]	0.23	0.002
Time Slope 2	2.29	[1.25, 3.33]	0.53	0.000
Time Slope 3	1.64	[0.23, 3.04]	0.72	0.022
SD (Intercept)	1.19	[1.03, 1.38]		
SD Time Slope 1	2.37			
SD Time Slope 2	4.72			
SD Time Slope 3	5.64			
Cor: Time Slope 1 and Intercept	-0.32			
Cor: Time Slope 2 and Intercept	0.36			
Cor: Time Slope 3 and Intercept	0.39			
SD (Observations)	1			

A visualization of the random effects is also presented in Figure 13. This graph shows individual cubic splines fit to each individual participant with two knots. While visually noisy, this graph conveys that there is a great deal of variability in individual trajectories throughout the study. Not only do the random intercepts vary across close to the entire y-axis, but the slopes also show a great deal of variability in the direction and steepness of their curves.

The ICC for this model was 0.51, which indicates that around half the variance in careless response propensity over time is accounted for by an individual's unique response patterns. The strong effect of the individual on careless response propensity is also highlighted in Figure 13, which shows that not all participants shared the general positive trajectory over time. Specifically, some individuals actually have negative, not positive, slopes for their final assessments, indicating that their probability of responding carelessly was decreasing, not increasing, towards the end of the study period. This is also represented in the random effects parameters in Table 5, which show a high

standard deviation for the second slope after the knot at time 10. The third slope also shows a high standard deviation, but again should be interpreted with caution due to the sparsity of the data.

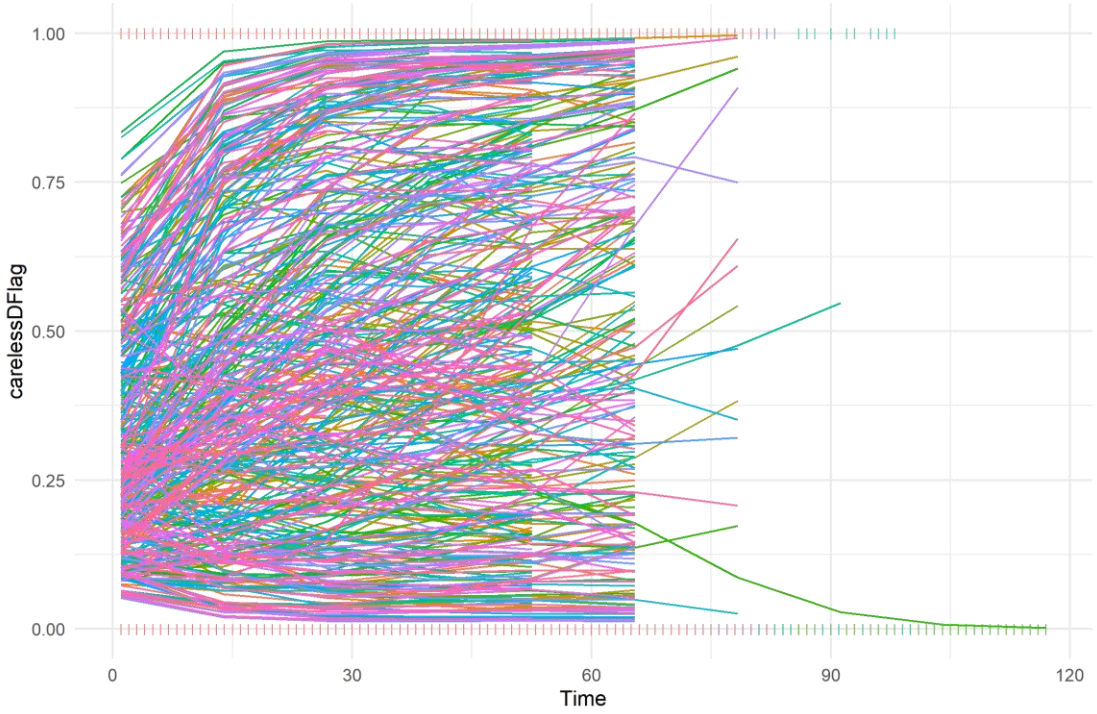

Figure 13. Careless Responding Over Time by Individual

Personality and Careless Responding Over Time

All personality variables were individually regressed onto the probability of being flagged as careless using cubic smooths to visualize the univariate relationships between variables. These graphs can be seen in Figure 14. All variables appear to have some relationship with the probability of being flagged as careless and theoretically these relationships make sense. Highly agreeable, conscientious, extraverted, and open people have a lower propensity to respond carelessly, whereas highly neurotic people have a higher propensity, though it is important to note that these graphs should not be interpreted too closely as they do not include random effects or time.

To keep results in line with the scope of this paper, only conscientiousness and agreeableness will be examined in detail. While this graph does highlight that other variables may have potential

relationships with careless response rates, examining these intricacies was determined to be beyond the scope of this paper. Similarly, only additive effects for personality were considered and not interactive effects. This choice was again made to keep results within the scope of this paper and because of the difficulties in fitting the originally proposed GAM models.

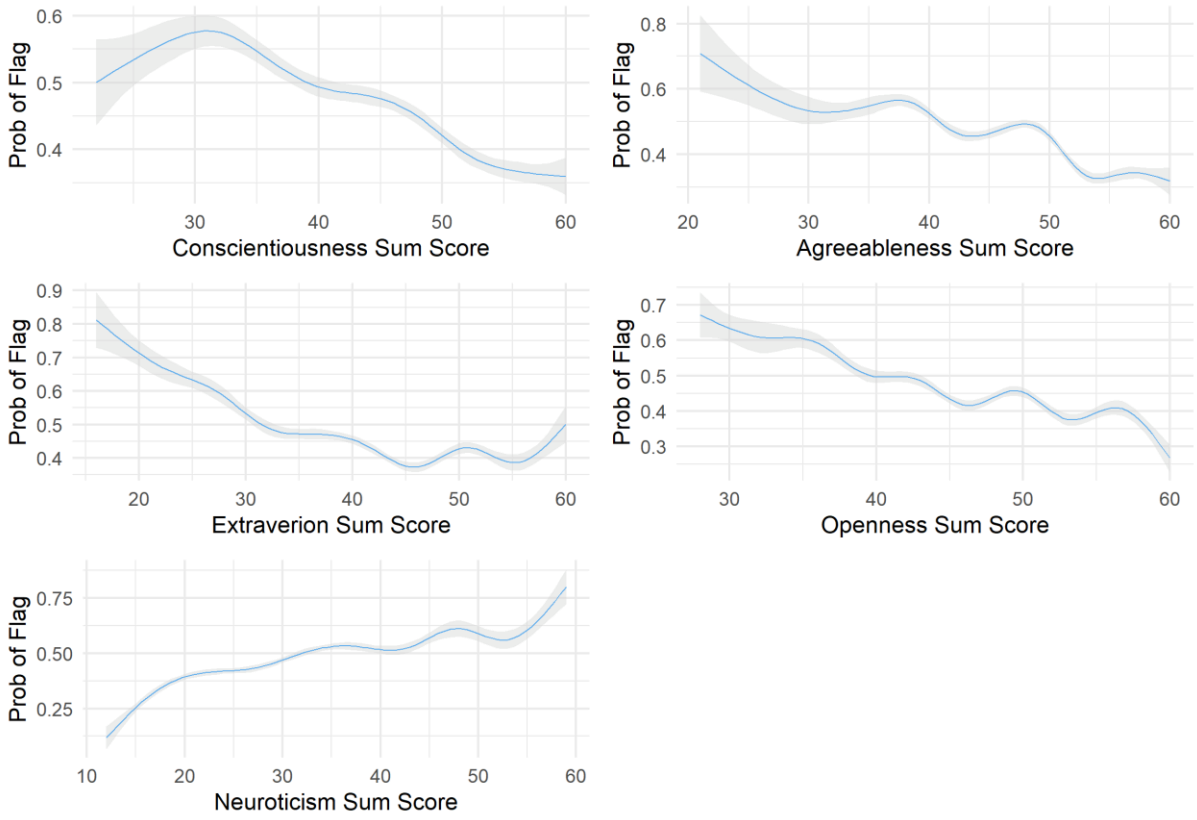

Figure 14. Careless Response Rates by Personality Score

An initial main effects model including conscientiousness and agreeableness was fit with a knot at 35 for conscientiousness and knots at 30 and 40 for agreeableness. The initial knot selection was guided by inspecting the above graphs. Inspecting the model parameters revealed that the estimate for the slope after the conscientiousness knot was significant as was the final slope for the agreeableness model. Confidence intervals ranged from -1.98 to 1.69 for the first conscientiousness slope and -6.43 to 3.28 for the second agreeableness slope. The reason for these wide intervals could

be due to suboptimal placement of the knots, and because of this placement was adjusted. Furthermore, the third knot for agreeableness was dropped, as the slope estimates for the second and third slope were nearly identical (1.48 vs 1.50) with only the wide confidence interval on the second slope differentiating them. AIC for the original model was 27551.446 and for a model with one knot at 30 on conscientiousness and 35 on agreeableness it was 27549.549. This represents a small improvement in information loss but given that the single knot model is simpler it was preferred.

Knots were moved to several points along the continuum for both traits and these models were compared. Preference was given to models with greater interpretability of the slopes for each knot. The final model chosen included a knot at 30 for conscientiousness and 35 for agreeableness. Slopes for both models were non-significant before the knot, but significant and negative after the knot. This, and the wide confidence intervals, suggest that there is not a consistent relationship between personality and careless responding for those who are low on these traits, but for individuals high on these traits the probability of responding carelessly is significantly reduced. In addition, slope estimates for time are nearly identical in this model compared to the model with no personality variables. This suggests that these variables capture unique variance in careless response propensity.

Again, because slopes were fit using cubic splines, parameter estimates do not have the same interpretation they would in a linear model. Figure 15 was created to visualize this model and the effects of personality on careless response rates. To aid in interpreting the visualization, subsets of data were taken at the mean (46.80 for agreeableness, 46.01 for conscientiousness) and standard deviation groupings. The standard deviation for both variables was ~8 and +1, -1, and -2 standard deviation groupings were created on both personality variables. The logic for including both -1 and -2 was to visualize changes close to the knots for both variables.

Table 6. Results for Personality Over Time Regression

Parameter	*b*	*b* 95% CI [LL, UU]	SE	p-value
(Intercept)	-0.13	[-1.75, 1.49]	0.83	0.87
Slope 1 Time	0.74	[0.30, 1.19]	0.23	0.001
Slope 2 Time	2.29	[1.25, 3.33]	0.53	1.53E-05
Slope 3 Time	1.61	[0.22, 3.01]	0.71	0.02
Slope 1 Conscientiousness	-0.02	[-1.96, 1.93]	0.99	0.98
Slope 2 Conscientiousness	-0.64	[-1.28, -0.01]	0.32	0.05
Slope 1 Agreeableness	-1.07	[-3.84, 1.70]	1.41	0.45
Slope 2 Agreeableness	-1.27	[-2.04, -0.50]	0.39	0.001
SD (Intercept)	1.17	[1.00, 1.36]		
SD Time Slope 1	2.35			
SD Time Slope 2	4.58			
SD Time Slope 3	5.45			
Cor Time Slope 1 and Intercept	-0.34			
Cor Time Slope 2 and Intercept	0.28			
Cor Time Slope 3 and Intercept	0.33			
SD (Observations)	1			

Figure 15 is paneled by agreeableness scores and lines are colored by conscientiousness scores. The overall trend of these relationships is that those higher on conscientiousness show a lower propensity to respond carelessly, as do those higher on agreeableness. The two negative standard deviation groups also show no differentiation from each other. Confidence intervals are not shown on

these graphs to avoid visual clutter, but keep in mind that slopes after time 60 have wide confidence intervals and should not be interpreted with too much detail.

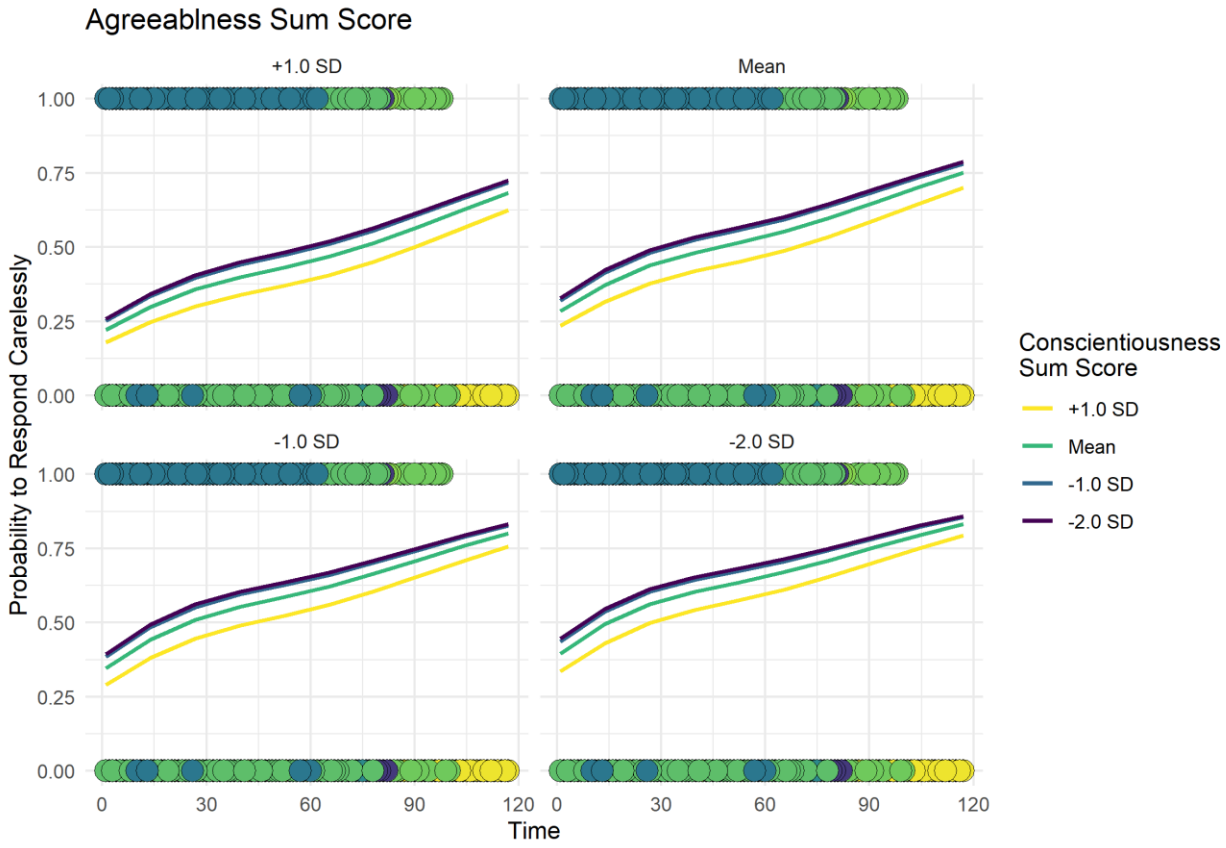

Figure 15. Careless Responding Over Time Paneled by Agreeableness and Grouped by Conscientiousness

Chapter 4: Discussion

Perhaps the most surprising finding from this study is the sheer amount of response episodes that were flagged as careless. Even using the most conservative cut scores, a quarter of response episodes were flagged, largely using the constituency indices. Identical responses to all items alone occurred in 10.51% of response episodes. One potential explanation is that the PANAS scale used naturally induces consistent responses due to items being similar to each other. However, theoretically, and empirically this seems unlikely. To take our criterion item of distressed-determined as an example, there is not an obvious explanation as to why these items would be related, nor does empirical evidence from the original validation study seem to support such a relationship as the correlation among these items was .06 (Kercher, 1992).

A second explanation could be that participants were often very low on the affect being assessed. That is, they rarely felt distressed, determined, enthusiastic, or excited and were consistent simply because of this lack of affect. While this certainly could be the case for some participants who were flagged, an examination of the various criterion graphs above reveals that participants were often flagged even when they did not respond with a "1". Thus, a simple lack of affect does not appear to explain this consistency in response patterns.

Work in the area of emotion differentiation, which is the degree to which individuals can differentiate emotional experiences, also offers an explanation for this consistency. This body of evidence finds that individual's psychosocial adjustment is a predictor of emotional differentiation and that depression, schizophrenia, alcohol problems, and other associated disorders are related to low levels of emotion differentiation (Smidt & Suvak, 2015). Because those low on emotional

differentiation have difficulty differentiating emotions, we might expect some individuals to have higher longstring values not because they are careless, but because they have difficulty differentiating emotional experiences. To examine whether this was the case, a Poisson multilevel model with random intercepts and slopes for time was fit to examine the relationship between longstring scores, time, baseline scores on neuroticism, and baseline scores on trait anxiety. Main effects terms were included for each variable and two-way interactions were also included for time with both baseline variables. The "HAC" sandwich estimator from the parameters package was used to calculate robust standard errors as the heteroskedasticity assumption showed evidence of violation. Results indicated that all slopes besides the main effect of time were centered firmly on zero. While trait anxiety and neuroticism are not ideal measures of psychosocial adjustment, this analysis does not find evidence that the factors most closely related to psychosocial adjustment in the current study relate to longstring.

To further examine whether this large degree of response consistency could be explained by the PANAS scale, the 1952 response episodes where the BFI10 was administered were examined. This scale assessed five constructs using ten items, so the above longstring cut score of six is not applicable. Furthermore, the constructs alternate within this scale, such that the first item measures extraversion, the second agreeableness, the third conscientiousness, etc. Thus, even a longstring of five on this scale is very unlikely, as that would represent identical responses across five different personality constructs. Of the 1952 response episodes, 5.79% had a longstring ≥ 5, 12.96%, had a longstring ≥ 4, and 37.14% had a longstring ≥ 3[10]. Eighteen response episodes, 0.92%, had a longstring of 10. While these estimates are certainly smaller than the proportion of PANAS episodes

[10] While a cut score of 3 may seem too liberal, remember that this represents identical response to three separate big five traits. Thus, this would represent someone who is exactly as agreeable as they are extraverted and conscientious.

flagged by longstring, there were still a large number of episodes flagged. This again suggestions that the response consistency observed on the PANAS is not simply due to the scale.

The reason for the decrease in BFI flags compared to PANAS flags could also be due to factors outside of the specific scale used. For example, an alternative explanation could be that participants were more accustomed to the PANAS measure because they saw it every day and items were always presented in the same order. Thus, quickly responding to these items in a way that seems largely plausible, in that it shows consistency, would be easier than responding in such a way to the following scales, which alternated each day. Future research could test whether participants have higher careless response rates if a scale is presented at every episode vs scales that alternate.

The response consistency findings also algin with those of Jaso et al., (2021) who found that participants exhibited a great deal of consistency. For example, in Figure 16 of their paper it is apparent that most respondents were flagged for having a SD $\leq$ 5 or a higher percentage of identical responses at the mode. It is difficult to directly compare the cut scores across our studies as the response scale used in their study was a 0-100 slider, however, there appears to be a great deal of agreement in that the most common type of response pattern was overly consistent responding.

It is somewhat unclear why overly consistent responding is so common, as general estimates in the careless response literature tend to find that inconsistent responding is more common. One explanation might be that when responding on a phone it is easier to consistently hit the same response option compared to moving one's finger randomly around the screen. Future research could further establish that this discrepancy exists by inducing participants to respond carelessly to the same scale using either a cell phone or computer and comparing the resulting response patterns.

Patterns and Predictors of Carelessness

The second major finding of this paper is that careless responding increases over time and is related to the theoretically relevant constructs of agreeableness and conscientiousness. While careless

responding showed a general increase over time, the pattern was not consistent for all individuals. Specifically, some individuals showed large increases in careless responding as the study progressed, while others showed a decrease in the probability of responding carelessly.

The results examining careless response episodes flagged within each individual and flagged after the first response episode also speak to this variability. The proportion of episodes flagged as careless after the first instance of carelessness ranged from 1.39% to 100% and 71% of participants had 25% or more of their response episodes flagged as careless. Because of the amount of variability in careless response behavior and suggests that a one size fits all approach may not be appropriate for careless response screening.

Recommendations for Screening for Carless Respondents

The findings from this study and from Jaso et al., both point to overly consistent responding as the most pressing concern for ESM studies. However, the cut scores determined in our studies highlight another issue, consistency indices are often scale and response option specific. As highlighted above with the BFI10, cut scores used on the PANAS may not be appropriate for other scales if the number of items measuring different constructs differs. Further, even the order of item presentation matters. In the present study all items were presented in a non-randomized order, however item randomization is quite common. In the case where items are fully randomized the researcher must determine if a lower longstring value might be appropriate given that a longsting of 5, for example, might no longer reflect consistent responses within a scale.

Further complicating this matter is the issue of response scale. As seen in the difference between cut scores in the present study and those of Jaso et al., response scale has a large impact on the selection of an appropriate cut score. Thus, researchers must think carefully about not only what they are measuring, but how they are measuring it in order to determine which cut scores may be appropriate.

Two metrics that suffer less from this challenge are response time and person total correlation. Response time, while still tied to the length of item content, behaves a manner that is easier to predict across measures and response scales. Similarly, since PTC is computed using the correlation scale it should remain relatively consistent across studies. However, it is worth noting that this method assumes that scales measure a construct that is unidimensional (e.g., ranging from low to high) and any scale violating this assumption may not produce interpretable results.

Finally, just as scale and response option selection are important so too is item selection. The present study contained no psychometric antonyms, which significantly hampered our ability to examine item correlations for overly consistent responding. While the distressed-determined pair acted as a workaround, ideally psychometric antonyms would also be available. Given that there is now growing evidence that overly consistent responding is the dominant response pattern in ESM studies, psychometric antonyms could prove to be a key tool to help researchers screen for these respondents while not compromising the content of the survey. These antonyms need not be reverse worded questions, but can be semantically opposite pairs such as extroverted-introverted, sad-happy, etc. These pairs can act as a validation check for any researchers attempting to develop cut scores in their own studies.

Finally, a decision must also be made about what to do with careless respondents once they are identified. In the current study, response episodes flagged as careless were not dealt with, as we did not address substantive research questions beyond the detection and predictors of careless response behavior itself. For most ESM studies, researchers are interested in substantive phenomenon unrelated to careless response behavior and this behavior simply acts as a nuisance in answering this question.

The results of our study indicate that there is likely not a black and white solution for dealing with careless respondents. For example, our results indicate that there are a subset of participants who responded carelessly to all of or nearly all of their response episodes. It is clear that these participants should be excluded from analysis because they provided data that was completely or almost completely invalid. However, what should be done for the participants who respond carelessly only a small number of times? Simply excluding these response episodes and treating them as missing data is the simplest approach, however this data is clearly not missing at random. Imputation also does not seem appropriate given that there is likely some external cause for their carelessness that day.

More nuanced approaches could be developed but require further research to understand why individuals respond carelessly only occasionally. Is there something in the external environment that distracts participants? Are participants simply feeling lazy or unmotivated? Or is there something else going on that is causing them to respond carelessly? One method for investigating this could be to compute careless response metrics shortly after the response episode occurs and send follow up pings to participants flagged as careless asking them if there is anything happening that might make them respond carelessly. The downside of this approach is obviously that participants may not be honest, but it could also lead to participants developing strategies to not getting caught when responding carelessly in the future, so should be used with caution.

A second problem when dealing with careless respondents is that not all carelessness is created equal. Take the 4.36% of response episodes that had identical responses within PANAS subscales but different responses between them. It seems highly unlikely that these respondents were truly giving a detailed report of their emotions at that point in time; however, it is also clear that these respondents are likely less careless than someone who responded identically to each question. Thus,

some form of downweighing might be appropriate, where responses that are somewhat careless are still included in analyses but given some weight to account for the fact they contain more error than a fully conscientious respondent.

As a last area of future research, simulation studies could provide a great deal of insight as to how best to do deal with careless respondents. Specifically, the effect of varying amounts of careless responding (e.g., 30% of episodes vs 70%) could be examined in addition to how different types of carelessness (e.g., fully content response vs partially responsive) effect study results. By doing so researchers could obtain a better understanding of how these different parameters effect statistical estimates in ESM studies and provide more detailed suggestions about how to deal with the various manifestations of careless responding in ESM studies.

Unfortunately, the answers to the above questions are likely that there is no simple answer and that a nuanced and multifaceted approach is required for dealing with careless responding in ESM studies. However, just because this problem is difficult does not mean that it is one ESM researchers can ignore. It is quite clear that careless responding is a problem, as even the most conservative estimate from this study places its prevalence at 25.4% of all response episodes. If ESM researchers wish to ensure that results from their study are not biased by careless responding they must think critically about how to detect and account for careless responding.

Overall, researchers could consider what likely and unlikely response patterns will look like in their data and build models that appropriately capture that unlikely data. This may require moving beyond simple cut scores. For example, while not examined in this study, a more nuanced consistency screen could ignore those who are low on all affect items, as these could be conscientious responses from people who are simply low on affect. While it is appealing to recommend a one size fits all approach, it seems unlikely that this will be effective at capturing

careless responding in all contexts. Similar to how statistical modeling is complex and context dependent, so too is careless response detection.

Despite these nuances, the steps followed in the current study can be applied more broadly and are as follows: 1) researchers should utilize all of the statistical methods discussed in this paper if possible; however, limitations such as single item scales may make ISD within impossible to compute, or response time may not have been tracked. Calculations for these indices are presented in Curran (2016) and are available in several R packages (Curran, 2018; Jaso et al., 2021; Yentes & Wilhelm, 2018).

Second, researchers can create an initial window of cut scores for each metric to examine potential careless responses. Creation of these cut scores should be guided by the properties of the scale and data simulation. For example, with short item stems such as the PANAS a window around 1 second per-item seems appropriate, whereas stems that include a sentence or two should be examined in the 2-3 second range. Longstring should be set at a theoretical minimum and examined in increasing increments from this minimum. For example, if a study involves assessing five constructs with three items each, an initial longstring cut of four could be used if it is expected that participations will change responses between constructs. In terms of ISD within, random uniform data can be generated using the *samples()* function in R or its analog in other programming languages to determine the standard deviation of random responses within constructs. Finally, the scales can be manually filled out using various patterned response sets and a standard deviation of these can be computed to assess possible values for ISD between. In most cases a PTC of 0 is a reasonable place to construct a cut window, but researchers should consider the possibility of rare but valid response patterns in their data. The standard deviation of these metrics can be used to set cut windows if the researchers cannot readily construct one. An initial window could be set using half or quarter

standard deviations on each metric and increased if this selects too few response episodes into each window. Regardless of the approach, these steps should be thoroughly documented and conveyed to the reader in the manuscript, appendices, or supplemental materials.

Third, after setting these windows, plots on criterion items can be generated and examined in much the same way they were in the current study. When selecting criterion items, researchers should choose items that have established statistical or semantic properties that can be used as the basis of comparison between groups. Ideally, psychometric synonyms and antonyms would both be examined; however, this may not always be possible. If this is the case, researchers could examine items that are expected to be uncorrelated, such as the positive-negative affect pairs used in this study. If neither psychometric synonyms or antonyms are available researchers should be aware that detecting inconsistent respondents may be difficult or impossible.

Fourth, after examining these criterion plots cut scores may need to be adjusted further in order to examine potential careless respondents. Researchers should document these modifications and provide original plots in supplemental materials for the paper. Once ideal cut scores are determined for each metric, careless respondents should be flagged and correlations between this subgroup and the overall sample should be compared. Researchers may decide to use more conservative or liberal cut scores in their analyses, but should provide justification for these choices.

Fifth, researchers should remove, downweight, or impute careless response data. In the case of an individual who is always or almost always careless, removal is likely the best option. For participants who are rarely or intermittently careless, downweighting or treating this data as MAR and imputing scores could both be used. Regardless, results should be reported both with and without careless data included, similar to common practice for outliers.

Chapter 5: Limitations

While providing a great deal of information about careless response behavior this study had several limitations. First, data for scales outside the PANAS and BFI10 were not available to the researchers for analysis. It is possible that some unknown features of the PANAS induced behavior that seemed careless but was in fact not. Further, it was not possible to disentangle specific aspects of the PANAS scale from potential effects of scale presentation frequency. That is, the high levels of carelessness on the PANAS could be representative, could be due to specific features of the PANAS, or could be due to the fact that it was presented every day. Future research could examine how different scales may or may not affect participant response styles and assess how the frequency of scale presentation moderates carelessness.

Second, examining personality variables outside agreeableness and conscientiousness was determined to be outside the scope of this study. Future research could examine how these, and other individual differences influence careless response trajectories over time. Additionally, the current study examined only additive effects of personality and not multiplicative. Future research could also examine how personality may interact with other variables to influence careless response rates.

Third, the lack of psychometric antonyms presents a major problem for validation of metric cut scores. Considering overly consistent responding seems to be the prevalent type of carelessness in ESM studies, psychometric antonyms seem crucial for proper validation of cut scores. Future research could not only investigate the above findings using psychometric antonyms, but test how many antonyms are required to produce accurate validation checks for careless response metrics.

Finally, while the current study illustrates that careless responding is clearly a problem, it does not allow for many conclusions about what can be done to fix this problem. Given that this was a working sample who was compensated well for their participation, simply providing competitive monetary compensation for study participation does not seem sufficient. Future research should examine both environmental and survey level factors that induce careless response behavior to allow researchers to create better surveys that may reduce careless responding. Until then, researchers are advised to proceed with caution when analyzing ESM data, as it appears many of the response episodes contained in this data could be contaminated by careless responses.

References

Baayen, H., & Linke, M. (2020). Generalized additive mixed models. In *A Practical Handbook of Corpus Linguistics* (1st ed., p. 686). Springer International Publishing.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beal, D. J. (2015). ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*(1), 383–407. https://doi.org/10.1146/annurev-orgpsych-032414-111335

Berkel, N. V., Ferreira, D., & Kostakos, V. (2017). The experience sampling method on mobile devices. *ACM Computing Surveys*, *50*(6), 1–40. https://doi.org/10/ggj74f

Brooks, M. E., Kristensen, K., Benthem, K. J. van, Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). GlmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

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

Curran, P. G. (2018). *CIE* [R]. https://github.com/paulgcurran/CIE

Curran, P. G., & Denison, A. J. (2019). Creating carelessness: A comparative analysis of common techniques for the simulation of careless responder data. *PsyArXiv*. https://doi.org/10/ggbgsq

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

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

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*(1), 105–113. https://doi.org/10.1177/001316446802800110

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*(1), 105–121. https://doi.org/10.1007/s10869-016-9479-0

Dupuis, M., Meier, E., & Cuneo, F. (2019). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods*, *51*(5), 2228–2237. https://doi.org/10.3758/s13428-018-1103-y

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*. https://doi.org/10.31234/osf.io/zf4nm

Gabriel, A. S., Podsakoff, N. P., Beal, D. J., Scott, B. A., Sonnentag, S., Trougakos, J. P., & Butts, M. M. (2019). Experience sampling methods: A discussion of critical trends and

considerations for scholarly advancement. *Organizational Research Methods*, *22*(4), 969–1006. https://doi.org/10/gfb6x5

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

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

Jaso, B. A., Kraus, N. I., & Heller, A. S. (2021). Identification of careless responding in ecological momentary assessment research: From posthoc analyses to real-time data monitoring. *Psychological Methods*. https://doi.org/10.1037/met0000312

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

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*. https://doi.org/10.1177/1094428115571894

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Kercher, K. (1992). Assessing subjective well-being in the old-old: The panas as a measure of orthogonal dimensions of positive and negative affect. *Research on Aging*, *14*(2), 131–168. https://doi.org/10.1177/0164027592142001

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*(53), 2445. https://doi.org/10.21105/joss.02445

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Lüdecke, D., Patil, I., Ben-Shachar, M. S., Wiernik, B. M., Waggoner, P., & Makowski, D. (2021). see: An R package for visualizing statistical models. *Journal of Open Source Software*, *6*(64), 3393. https://doi.org/10.21105/joss.03393

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/gddnqh

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology*, *65*(2), 287–321. https://doi.org/10.1111/apps.12058

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computin. https://www.R-project.org

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Schneider, S., May, M., & Stone, A. A. (2018). Careless responding in internet-based quality of life assessments. *Quality of Life Research*, *27*(4), 1077–1088. https://doi.org/10.1007/s11136-017-1767-2

Smidt, K. E., & Suvak, M. K. (2015). A brief, but nuanced, review of emotional granularity and emotion differentiation research. *Current Opinion in Psychology*, *3*, 48–51. https://doi.org/10.1016/j.copsyc.2015.02.007

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Spielberger, D. (1983). *State-trait anxiety inventory for adults* [Inventory]. https://journal.sipsych.org/index.php/IJP/article/view/620

van de Mortel, T. (2008). Faking it: Social desirability response bias in selfreport research. *Australian Journal of Advanced Nursing*, *25*(4).

Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form* (Department of Psychology Publications) [Manual]. Iowa Research Online. https://doi.org/10.17077/48vt-m4t2

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*(2), 105–121. https://doi.org/10/c4mn8t

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wiernik, B. M., Ones, D. S., Marlin, B. M., Giordano, C., Dilchert, S., Mercado, B. K., Stanek, K. C., Birkland, A., Wang, Y., Ellis, B., Yazar, Y., Kostal, J. W., Kumar, S., Hnat, T., Ertin, E., Sano, A., Ganesan, D. K., Choudhoury, T., & al'Absi, M. (2020). Using mobile sensors to study personality dynamics. *European Journal of Psychological Assessment*, *36*(6), 935–947. https://doi.org/10.1027/1015-5759/a000576

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986

Yentes, D., & Wilhelm, F. (2018). *careless: Procedures for computing indices of careless responding* (1.2) [R package]. https://github.com/ryentes/careless

Appendix A: R Models Run to Predict Carless Responding

Model	Code
GAM Time Model	bam(carelessFlag ~ s(Time, k = 40) + s(Time, ParticipantID, bs = "re"), family = "binomial")
GAM Time and Personality Model	bam(carelessFlag ~ ti(Time, k = 40) + ti(BFI2CSum) + ti(BFI2ASum) + ti(Time, ParticipantID, bs = "re"), family = "binomial")
Linear Time Model	glmmTMB(carelessFlag ~ Time + (Time|ParticipantID), family=binomial)
Cubic Spline Time Model	glmmTMB(carelessFlag ~ ns(Time, df = 3, knots = c(15,60)) + (ns(Time, df = 2, knots = c(15,60))|ParticipantID) family=binomial)
Cubic Spline Time and Personality Model	glmmTMB(carelessFlag ~ ns(Time, df = 3, knots = c(10,60)) + ns(BFI2CSum, df = 2, knots = 30) + ns(BFI2ASum, df = 2, knots = 35) + (ns(Time, df = 3, knots = c(10,60))|ParticipantID), family=binomial)