Psychological Methods

Identification of Careless Responding in Ecological Momentary Assessment Research: From Posthoc Analyses to Real-Time Data Monitoring

Brittany A. Jaso, Noah I. Kraus, and Aaron S. Heller

Online First Publication, September 16, 2021. http://dx.doi.org/10.1037/met0000312

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000312

Identification of Careless Responding in Ecological Momentary Assessment Research: From Posthoc Analyses to Real-Time Data Monitoring

Brittany A. Jaso, Noah I. Kraus, and Aaron S. Heller
Department of Psychology, University of Miami

Abstract

With the emerging ubiquity of cell phones, ecological momentary assessment (EMA) as a set of methods enable researchers to study momentary social, psychological, and affective responses to everyday life. Additionally, EMA enables researchers to acquire longitudinal data without the need for multiple lab visits. As the use of EMA in research increases, so too does the necessity of determining what constitutes valid or careless individual EMA responses to ensure validity and replicability of findings. Currently, EMA studies solely consider the response rate of a participant for exclusion. Yet, other features of an assessment can help to determine whether a response is careless or implausible. Here, we examined over 18,000 EMA text message responses of individual affect items to derive a data-driven model of what constitutes a "careless response." Results from this study indicate that an overly fast time to complete items ($\leq$1 s), an overly narrow within assessment response variance ($SD \leq 5$), and the percentage of items that fall at the mode ($\geq$60%) are independent and reliable indicators of a careless response. Excluding careless responses such as these remove implausible positive correlations among psychometric antonyms (e.g., relaxed and anxious). Further, by identifying and removing careless responses, we also identify careless responders, participants who could be removed from group analyses. We use these results to develop and introduce an R package, EMAeval, so EMA researchers may similarly identify careless responses and responders either online during data collection or posthoc, after data collection has completed.

Translational Abstract

Using mobile technology to sample people's experiences as they go about their daily lives has quickly become a central method in Psychology research. These methods have allowed psychologists to better understand emotions and cognitions as they are experienced in daily life. These methods also allow psychologists to better understand how individuals with psychiatric disorders differ in their daily emotional experiences from those without any psychopathology. While this research area has blossomed in recent years, there remains no standardized approach to know if an assessment of real-world emotion has been completed thoughtfully or carelessly. Ensuring the quality of our real-world data is critical as not doing so may impact the conclusions researchers make. Here, we examine over 18,000 assessments of emotional experience collected from cell phone surveys and develop and test metrics that all researchers who rely on cell phone surveys can use to identify a response as potentially invalid, or careless. These three metrics include: (a) how quickly an assessment has been completed (where completing too quickly is likely invalid); (b) whether the responses are within a restricted range or cover a broader range (where too restricted a range is likely invalid); and (c) if the responses within the assessment are mostly identical. We present an R-package (https://github.com/manateelab/EMAeval-R-Package) for the research community to apply these criteria to their own work to enhance the validity of their experience sampling data.

Keywords: ecological momentary assessment, response quality, real-time data monitoring, careless responses, R package

Brittany A. Jaso https://orcid.org/0000-0002-3474-5639
Aaron S. Heller https://orcid.org/0000-0002-0680-3248

The R-functions are downloadable at the lab's github page: https://github.com/manateelab/ preliminary findings from this article were accepted to the Society for Affective Sciences conference but were not disseminated due to the COVID-19 outbreak.

Correspondence concerning this article should be addressed to Aaron S. Heller, Department of Psychology, University of Miami, P.O. Box 248185, Coral Gables, FL 33124-0751, United States. Email: aheller@miami.edu

With the emerging ubiquity of cell phones, researchers are rapidly gaining insight into the daily lives of research participants. Ecological momentary assessment (EMA) as a set of methods enable researchers to study momentary social, psychological, and physiological responses to everyday life (Bolger et al., 2003; Bolger & Zuckerman, 1995; Suls et al., 1994; Wheeler & Reis, 1991). This is because, in developed nations, cellphones are omnipresent, and as such, these measurement devices are always with us and able to actively or passively monitor behaviors. As of 2019; 81% of Americans reported owning a smartphone (Anderson, 2019). Among smartphone owners, 58% of 18- to 29-year-olds reported they use their smartphone as their primary tool to access the Internet (Anderson, 2019). As technological advancements continue, the use of, and reliance on smartphone technology will too (Anderson, 2019), making smartphones an excellent tool to investigate real-world affect, cognition, and behavior.

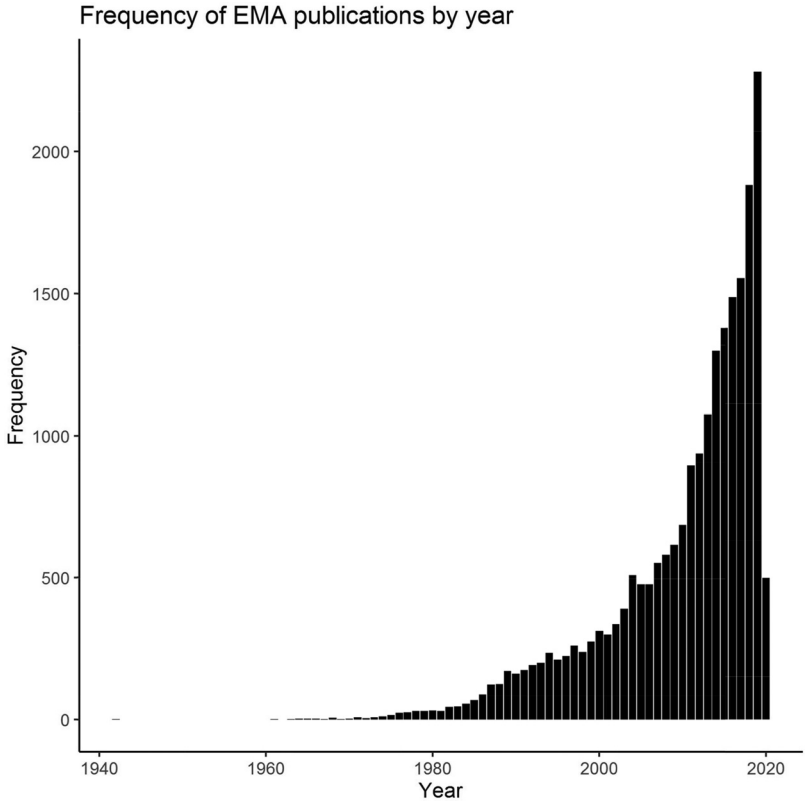
There are multiple advantages to cell-phone based EMA designs, which is why studies using EMA have become popular in recent years (Van Berkel et al., 2018; Figure 1). One advantage to EMA methods are that surveys assessing affect, cognition, or behavior, can do so currently, "in the moment," rather than retrospectively asking participants during a lab visit to report on their previous week or weeks. By measuring experiences as they occur in real time, EMA researchers need not be concerned with biases associated with retrospective self-reporting (Beal & Weiss, 2003; Schwarz & Clore, 1983). An additional attribute unique to modern EMA (as compared with older diary studies) is the presence of timestamps indicating precisely when participants begin and complete assessments. These timestamps help ensure participants are not "backfiling" or completing their assessments at the conclusion of the study (Hufford, 2007; Wen et al., 2017).

A second advantage of EMA designs is that it is relatively easy to acquire data longitudinally. Longitudinal studies allow for a closer examination of between- and within-person processes (emotions, cognitions, and behaviors). These methods enable researchers to map out the evolution of variable relationships as they emerge across time, compared to brief cross-sectional snippets (Bolger & Laurenceau, 2013; Molenaar et al., 2003; Wang et al., 2012). An additional benefit to in-the-moment, longitudinal designs is that researchers can assess low-intensity and low-frequency behaviors that, when asked about retrospectively, may be forgotten (Bolger et al., 2003; Bolger & Laurenceau, 2013).

A more understated benefit of EMA studies relates to the research participant experience. Given the ubiquity of cell phones, it may be less disruptive for research participants to receive text messages or cell phone notifications compared with attending multiple in-person lab visits. A review article that estimated attrition rates across 27 EMA studies in elderly patients, a population that

Figure 1

Frequency of Publications That Reference "Ecological Momentary Assessment" or "Experience Sampling" on Pubmed



Note. This image generated using Pubmed data up until February 2020.

tends to be less technologically savvy, reported that attrition rates were approximately 25% (Cain et al., 2009). While high, this attrition rate is comparable with lab-based longitudinal studies in which attrition rates are around 21% overall (de Graaf et al., 2000), 32% among participants ages 18–23 (Young et al., 2006), and 16% among older populations (Young et al., 2006). Even with the relatively high attrition rate reported in the Cain et al. review, it was suggested that attrition was attributed to difficulty understanding the instructions, not participant burden. In fact, 80% of participants reported that the EMA experience was not aversive (Cain et al., 2009). Further, in an investigation into the effects of study duration on response compliance, EMA response rates were similar for studies lasting >3 weeks compared EMA studies lasting <1 week (Wen et al., 2017). As such, the use of EMA for gathering data is not unduly burdensome.

Determining "Good Data" in EMA Studies

As the use of EMA in research studies increases, so too does the importance of determining what defines a valid EMA response and what defines a "compliant" research participant. Despite the statistical and theoretical benefits to capturing longitudinal data as individuals live their daily lives, there are currently no standard practices to clean EMA data. The closest "best-practice" standard is the tendency to remove participants who do not meet an a-priori compliance cut off defined by the percent of surveys completed. In the literature, there is a general trend to only include participants with a 70%–90% compliance rate, although this is selected by convention and not empirically determined (Beal & Weiss, 2003; Csikszentmihalyi, 2006; Csikszentmihalyi & Larson, 2014; Csikszentmihalyi & Larson, 1992; Ebner-Priemer et al., 2007; Ebner-Priemer & Trull, 2009; Jones et al., 2019; Vansimaeys et al., 2017).

While response rate cutoffs are necessary to identify noncompliant participants, there is almost no research addressing the "best practices" for determining the quality of individual EMA responses. The identification of careless responses, even in individuals who complete the majority of EMA surveys, may be of even greater importance to data validity than simply participant compliance (Hufford, 2007; McCabe et al., 2012). It is imperative that the data collected are cleaned to maximize the reliability, power, and validity of results. Because researchers using EMA are often interested in interpreting within-person effects and individuals differences, without proper data cleaning procedures, careless responses could be incorrectly attributed to individual differences (Beal & Weiss, 2003). We contend that the nascent EMA field would benefit from standardized methods to identify both careless responses and careless responders.

Existing Methods to Identify Careless Responses

Classic, cross-sectional survey research has identified several methods to identify careless responses (Meade & Craig, 2012). One method involves inserting special items that specifically test participants' attention (e.g., "If you are reading this, please select Option 3"). These sanity-checks are useful in that they flag participants who are not carefully reading items or not paying attention to the survey or task. Such methods are typically employed in the midst of lengthy, single-administration surveys (Meade & Craig, 2012). Yet, because EMA studies tend to assess identical constructs (often by asking the same questions) repeatedly over time, inserting sanity-check items may not be the most beneficial use of limited space and time. If the inclusion of special items is not a viable option due to concerns surrounding participant burden, alternative approaches to identifying careless responses and responders are needed.

Other methods to identify careless responses typically include post hoc analyses of data once surveys are completed. Although there is no consensus for how to clean survey data once data collection is complete, one method for determining the quality of responses is to examine the response time, with the hypothesis that overly fast responses or responders may be careless. However, there is no clear standard for what defines a survey that has been completed too quickly, because questions can vary in length or complexity (McCabe et al., 2012). Approaches to identify overly fast responses can be taken from the experimental psychology literature where reaction time (RT) is assessed to ensure responses were not completed randomly, or "by chance." However, task-based reaction times are often analyzed on the scale of milliseconds (e.g., anything quicker than 30 ms), which is too fast for survey questions (Christensen et al., 2003). Other researchers recommend using visual inspection of the distribution of response times to identify clear cut-offs for the determination of a "too quick" or "careless" response (McCabe et al., 2012). While the methods to identify quick responses vary across analysis and study design, there is consensus that overly fast responding is an indication of carelessness and should be considered for removal.

Other methods for identifying careless responses suggest investigating the actual responses themselves, as well as their covariance. If almost all responses in a single survey are given the same value (a high response pattern), it is likely that the report is not valid (McCabe et al., 2012). Termed "longstring," this method calculates the number of items or percent of the survey that received the same, modal value (Johnson, 2005). Longstring is a useful method for researchers to flag responses or participants who respond similarly to most survey items, especially if items assess psychometric antonyms (e.g., reverse scored items, inversely

Table 1
Demographic Breakdown of the Overall Sample and Separate Cohorts

Variable	2,000	3,000	5,000	6,000	8,000	9,000	Total
Number of participants	78	31	71	53	24	36	293
Age *M* (*SD*)	18.88 (2.6)	18.88 (1.11)	19.34 (2.5)	19.31 (1.5)	19.57 (1.36)	19.07 (0.56)	19.12 (2.46)
Sex% Female	71.05%	54.84%	75%	61.82%	50%	81.08%	68.26%
Race% Caucasian	72.37%	51.61%	70.42%	73.58%	70.83%	78.38%	70.65%
Race% African-American	11.84%	16.13%	12.5%	7.19%	8.33%	10.81%	11.26%
Race% Other	15.79%	32.26%	17.08%	19.23%	20.84%	10.81%	18.09%

related constructs). Psychometric antonyms are items that theoretically or logically should have a large difference (e.g., wakefulness and sleepiness; Goldberg, 2001; Johnson, 2005; Meade & Craig, 2012). Thus, in addition to examining time-stamps for speed of response, looking at the relationships between items can be valuable in ascertaining careless responses. Despite direction for how to clean data from classic survey literature, these methods have not been applied to identify careless patterns of EMA responding.

Current Study

Currently, there is no standardized method or freely available statistical package to identify careless EMA responses that should be flagged for removal. The purpose of the current study is to use a data-driven approach to examine over 18,000 EMA responses from 293 subjects in order to identify careless responses and responders. Using this analysis, we then develop a hierarchical system that will flag careless responses in real-time. Based on the results from our investigation, we introduce an R-package, EMAEval (available from our lab's github page: https://github.com/manateelab/EMAeval-R-Package, or via (devtools::install_github ("manateelab/EMAeval-R-Package"))), that will notify researchers in real-time when participants' responses are deemed careless so researchers can notify participants as needed.

Method

Participants

Participants included 293 undergraduate students enrolled in introduction to psychology and introduction to chemistry courses at the University of Miami. The sample consisted of seven cohorts, collected across seven semesters from Fall 2016 until Spring 2019; with a total of 18,093 assessments (*M* number of assessments per participant = 49). Ages ranged from 17–40 (*M* = 19.12

Table 2
The Breakdown of Emotion Items Assessed Within Each Cohort

Emotion Item	2000s	3000s	5000s	6000s	8000s	9000s
Happy						
Excited						
Relaxed						
Content						
Attentive						
Tired						
Upset						
Irritable						
Nervous						
Sluggish						
Anxious						
Sad						
Stressed						
Sleep (Quality)						
Sleep (Hours)						
Number of Items	11	11	10	10	12	12

KEY
Included
Not Included

Note. In tables and figures, the following cohorts are grouped as they completed identical item sets: (1) 2,000 and 3,000, (2) 5,000 and 6,000, (3) 8,000 and 9,000. The total number of items per assessment is included at the bottom of the table. See the online article for the color version of this table.

years, *SD* = 2.46), all participants under the age of 18 provided parental consent. Two-hundred (68.26%) participants identified as female, and 207 (70.65%) participants identified as Caucasian/White. Participants were compensated with course credit upon completion of the study. The breakdown of participant age, sex, and race by cohort can be found in Table 1.

Selected Items From the Positive and Negative Affect Schedule—Expanded Form (PANAS-X)

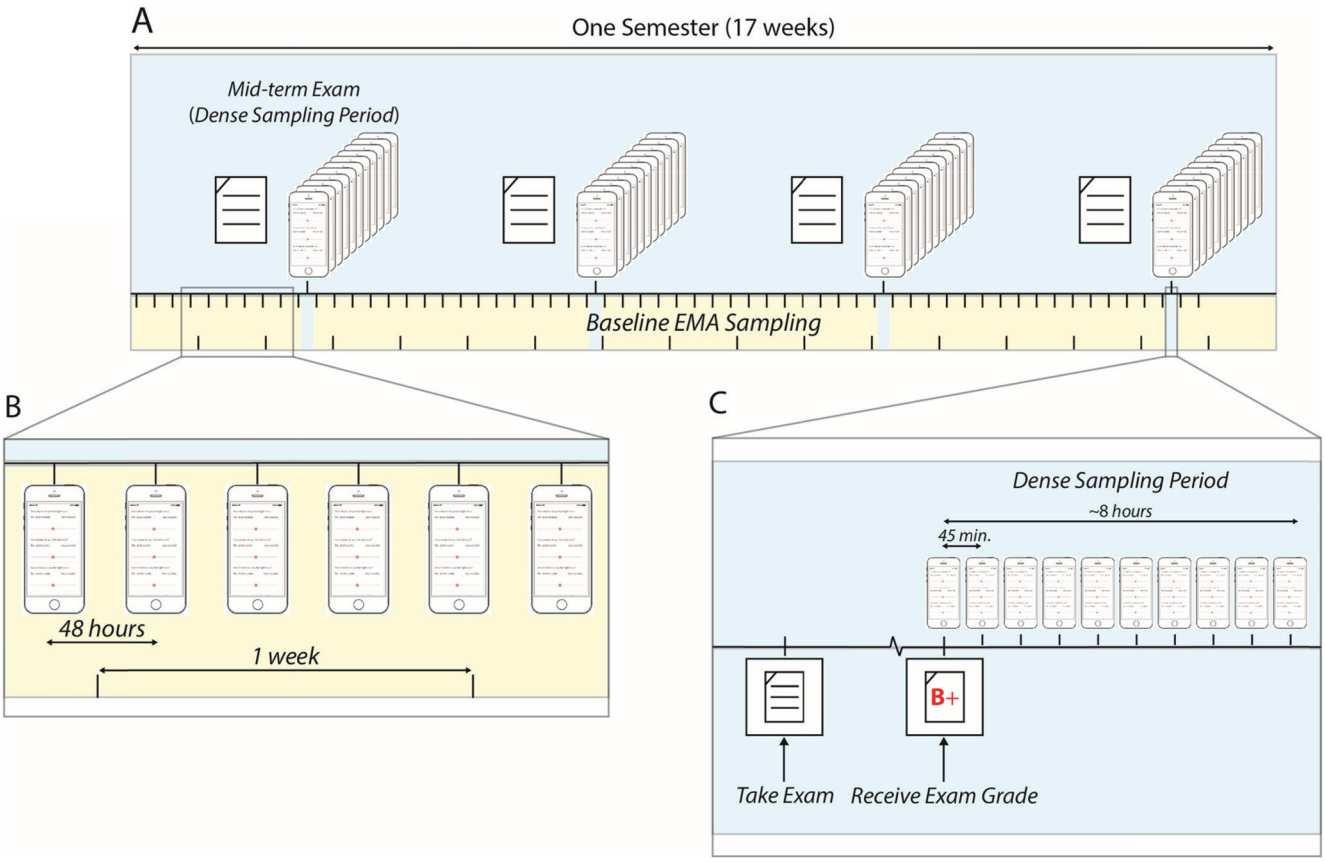
The PANAS-X (Watson & Clark, 1999) is a self-report measure, which assesses current levels of positive and negative emotion. The scale asks participants to rate how much they feel 60 different emotions in the present moment on a 5-point Likert scale from 1 (*very slightly or not at all*) to 5 (*extremely*). The PANAS-X has demonstrated evidence of good internal consistency (PA: α = .83–.9, NA: α = .85–.9) as well as convergent (r = .85–.95) and discriminant (r = −.02––.18) validity (Watson & Clark, 1999). These reliability and validity values are similar to the original, 20-item PANAS (Watson et al., 1988). The current study selected items from the original PANAS-X scale to satisfy the research purposes. In order to reduce participant burden, we chose to use up to 12 items, though the specific items varied somewhat depending on the cohort. A breakdown of PANAS-X items by cohort is available in Table 2. The 5-point Likert scale used to report participant response options was adjusted to a visual analog scale (VAS) that ranged from 0 (*not at all*) to 100 (*very*). The marker for the VAS was situated at the 50 so that participants "started" in the middle of the range and could move the marker to the left or right according to the strength of their emotion. This enabled researchers to investigate a greater range of intensity for each emotion item.

Data Collection

Study Procedure

The study was conducted for the duration of the academic semester following participants' consent (an average of 4 months). Interested participants were brought in for an initial lab visit. During the lab visit, participants provided informed

Figure 2
Study Design



Note. (A) Ecological momentary assessment (EMA) data were collected during a 17 week academic semester. Data collection included both "non-dense" baseline EMA sampling (B) in which EMA prompts were sent every 48 hr; and a "dense sampling period (C) which occurred immediately following exam feedback, in which EMA prompts were sent every 45 min for approximately 8 hr. Figure modified from (Villano et al., 2020). See the online article for the color version of this figure.

consent, contact information (i.e., cell phone number, email address) for purposes of EMA survey distribution and received instructions for how to complete the study. The study protocol was approved by the Institutional Review Board at the University of Miami.

EMA Self-Report Surveys: Sampling Period and Sampling Rate

As part of a larger study, participants received SMS messages at two different sampling rates.

Sampling Rate 1: "Nondense Sampling". For nondense (i.e., nonburst) data collection, participants received an SMS text message once every 2 days, which was randomly transmitted within an 8-hr time frame between the hours of 10 a.m. and 6 p.m. The text messages contained a brief sentence asking the participant to "Complete the survey below" with a personalized Qualtrics link. The messages were generated using in-house built software using a FileMaker platform connected with Twilio to automate the schedule and sending of SMS messages. Participants were instructed to respond to this survey within 4 hr of receipt.

Sampling Rate 2: "Dense Sampling". In addition to the nondense EMAs, participants were also sent survey links more frequently (every 45 min for 8.25 hr equaling out to approximately 11 surveys per dense sampling period) following the release and viewing of their exam grades (Villano et al., 2020). On average, there were four exams per semester, and thus students could participate in up to four dense-sampling periods throughout the duration of the study. Students were instructed to complete the survey as soon as they received it. Responses were automatically collected and stored on the Qualtrics website and downloaded after study completion. For a visual depiction of the EMA sampling periods, see Figure 2.

Data Analyses

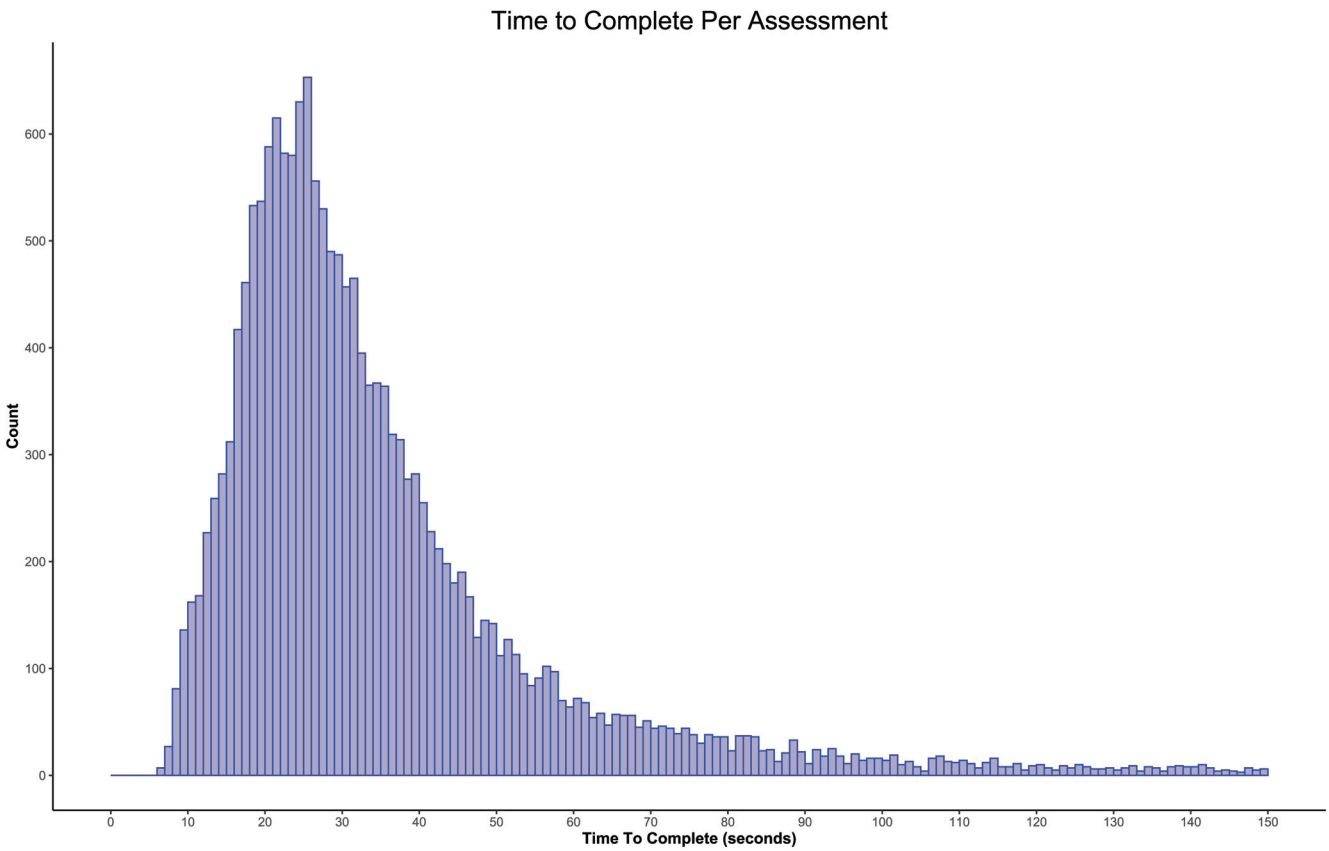
EMA responses from the seven cohorts were compiled into a single data-file. Based upon previous work suggesting specific indicators of survey responses may reflect a careless response (Goldberg, 2001; Johnson, 2005; Meade & Craig, 2012), we calculated and examined the following variables:

Time to Complete (TTC) per Assessment

Time to complete was operationalized as the difference between when the participant completed the EMA survey (submitted the Qualtrics survey) and when the participant started the

Figure 3

Distribution of Time to Complete (TTC) per Individual Assessments Across All Cohorts



Note. X-axis is truncated to assessments with a TTC ≤ 150 s to better visualize the distribution. To do so, this histogram excluded 653 assessments (3.61%) whose TTC > 150 s. See the online article for the color version of this figure.

EMA survey (clicked on the Qualtrics link to open it; end time—start time). For a visual depiction of the TTC distribution for our sample, see Figure 3. For a visual depiction of TTC by cohort, see Figure 4A.

Time per Item (TPI) per Assessment

Time per item was operationalized as the quotient of the individual EMA's TTC and the number of items in the EMA assessment. The TPI was defined in units of seconds. For a visual depiction of TPI by cohort and sample (see Figures 4B and 5, respectively). For analyses using TPI, we examine four separate windows: TPI $\leq$ 1 s; 1 $<$ TPI $\leq$ 1.5 s; 1.5 $<$ TPI $\leq$ 2 s; 2 $<$ TPI $\leq$ 3 s. We examined each window separately so the quality of responses at this specific TPI can be independently assessed. In addition, to evaluate whether particularly long response times also reflected a careless response, we also examined slow TPIs using 10 s and 20 s as cutoffs.

While TTC and TPI are obviously interrelated, because studies use different numbers of items per EMA, we perform most analyses using TPI and not TTC. Analyses are done with TPI so that other researchers can apply our findings to their research more easily by accounting for the number of items.

Mean, Standard Deviation, and Longstring of Emotion Items

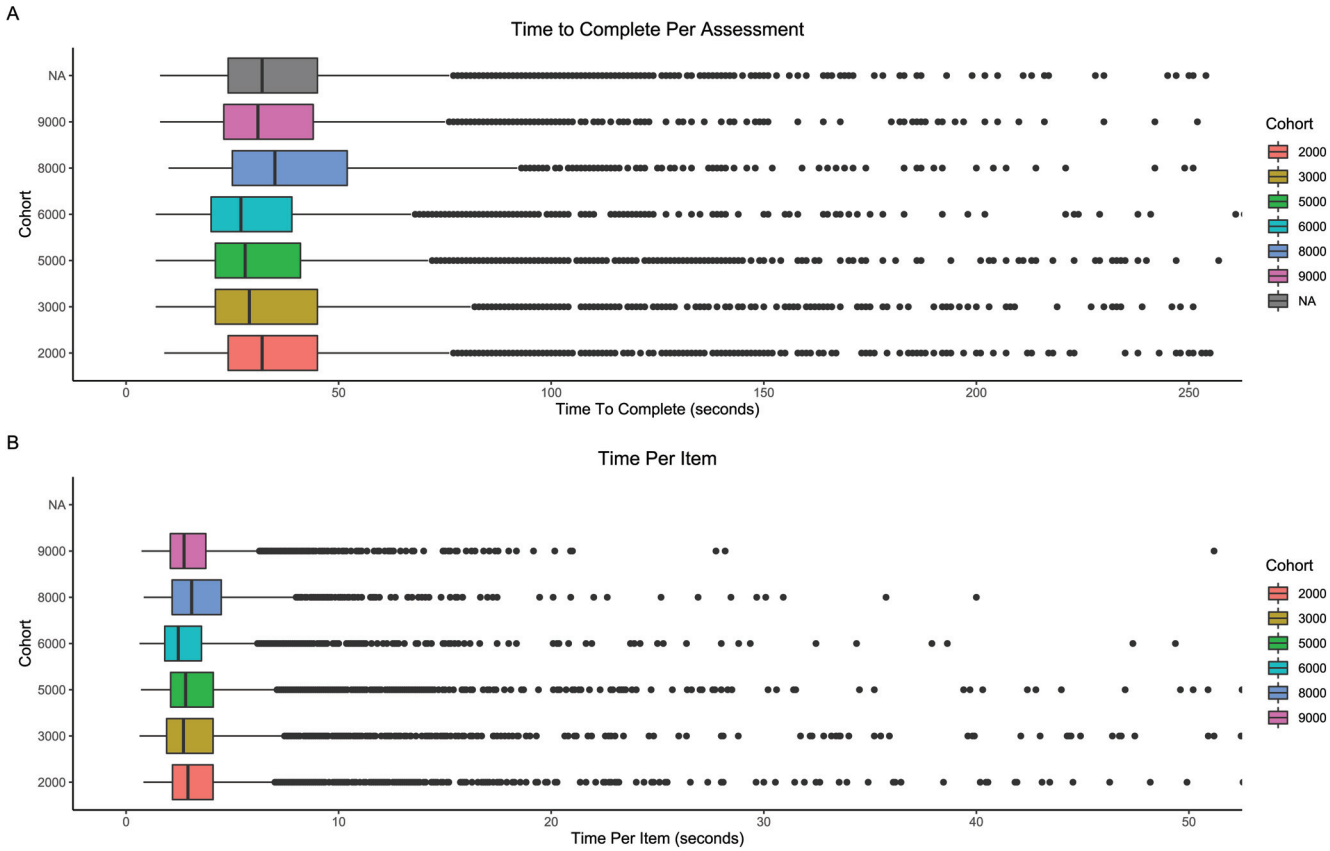
The mean, standard deviation (*SD*; Figure 6), and longstring (Figure 7) across items (given on a scale from 0 = *not at all* to 100 = *very*) was calculated for each EMA response. Similar to the TPI analysis, we used independent windows to assess the quality and plausibility of a response based upon within EMA *SD*. The windows included, $SD \leq 1$; $1 < SD \leq 2$; $2 < SD \leq 5$; $5 < SD \leq 10$; $10 < SD \leq 20$; $20 < SD \leq 30$; $30 < SD \leq 50$; $SD > 50$. Lastly, the longstring value was operationalized as the mode of the item scores per EMA (see Figure 7) and we calculated the percent of items at the mode for each EMA for levels of 50%, 60%, and 70% of items at the mode.

Correlation of EMA Items

Correlations between all item pairings were calculated (see Table 3). Because anxious and relaxed are psychometric antonyms in that they differ in both valence and activation (Feldman Barrett & Russell, 1998; Russell, 1979, 1980), and should, in general, have an inverse relationship, we examined whether the relationship between the rating of anxious and relaxed shifted with different careless response metrics and criteria (TPI, within EMA *SD*, % of items at the mode).

Figure 4

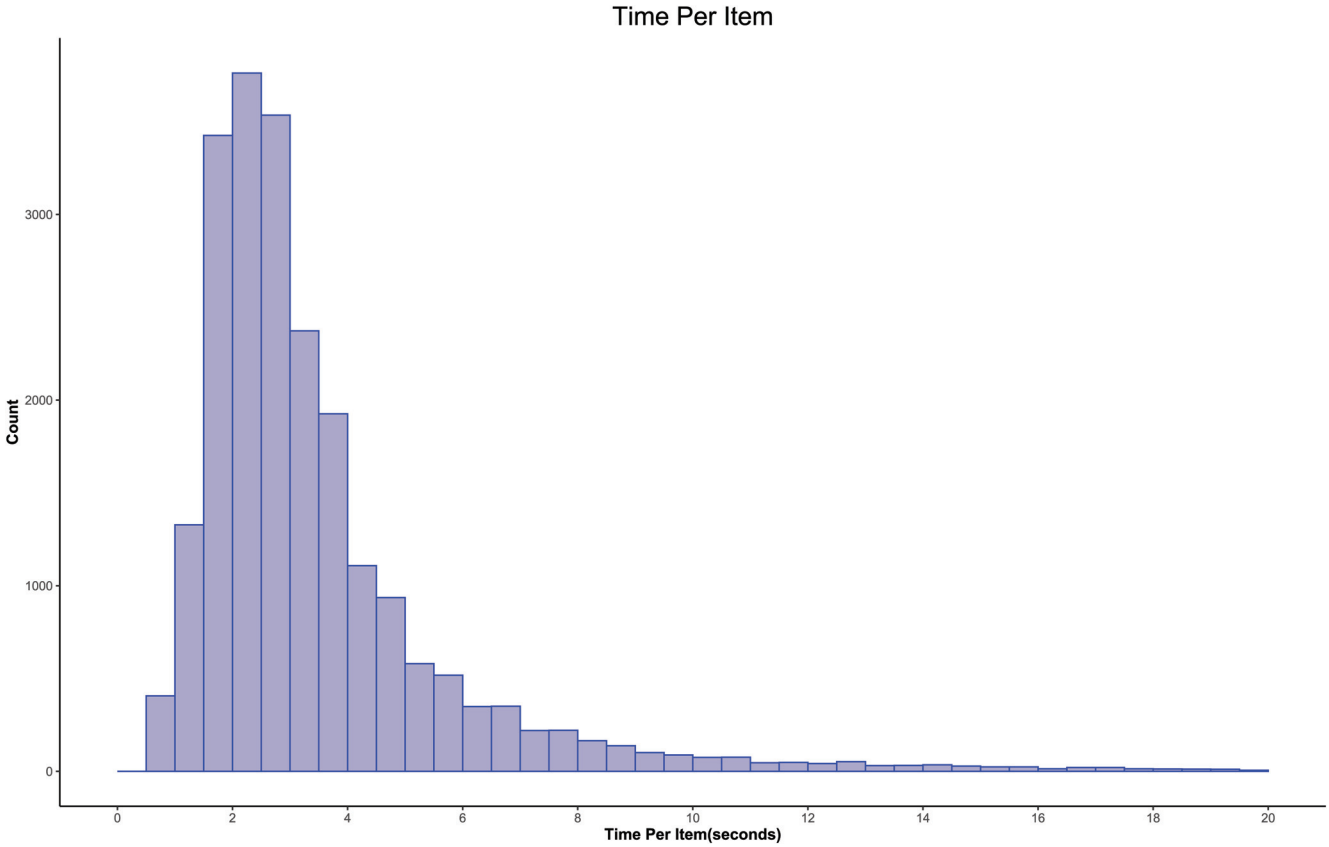
Boxplots Depicting the Distribution of the Response Data by Cohort



Note. (A) Distribution of time to complete (TTC) per assessment for each cohort. (B) Distribution of time per Item (TPI) for each cohort. See the online article for the color version of this figure.

Figure 5

Distribution of the Time per Item (TPI) for Individual Assessments Across All Cohorts



Note. X-axis is truncated to assessments with a TPI ≤ 20 s in order to better visualize the distribution. 355 assessments (1.96%) were excluded as they exceeded 20 s. See the online article for the color version of this figure.

Group z-Score Analysis

We additionally determined whether the number of EMA items per assessment (which differed slightly per cohort) affected the TTC and TPI distribution. If the number of items did affect the TTC/TPI distribution, this would suggest that the TTC or TPI cutoff would be less generalizable and more study-specific. To explore this question, we grouped cohorts by the number of items per EMA and then calculated z-scores for TTC (see Figure 8) and TPI (see Figure 9).

Quartile Removal Process

Given that the TTC for all EMA data were positively skewed (skew = 70.56; Figure 3) with some extreme outliers ($z > 4$), we first applied the upper fence of the interquartile rule to identify and remove EMAs whose TTC were excessively long (Stamatis, 2002). This interquartile rule identifies problematic outliers if they fall outside of the upper fence using the formulas below, where IQR stands for the interquartile range:

$$\text{Upper Fence} = Q3 + 1.5 \times IQR$$

Thus, the upper interquartile fence was used to first identify and remove EMAs that were not completed in a plausible timeframe per study recommendations. Once these extreme TTC EMAs were removed, additional metrics could be examined to identify careless EMAs.

Estimating How Many EMAs Are Necessary for a Stable Estimate of One's Affect

We estimated the number of EMAs required for a stable estimate of one's mean affect and affective variability. For each subject, we randomly sampled a subset of their EMA data (1, 2, . . . , 35 EMAs) and calculated the mean/*SD* of positive affect (PA) for that random sample. We also calculated the mean/*SD* of all EMAs for each subject as their "ground truth" estimate of mean PA and PA *SD*. We then calculated the across subject's rank-order correlation between the ground truth and the random sample. We performed this procedure four times to visualize the stability of the rank-order correlations.

Results

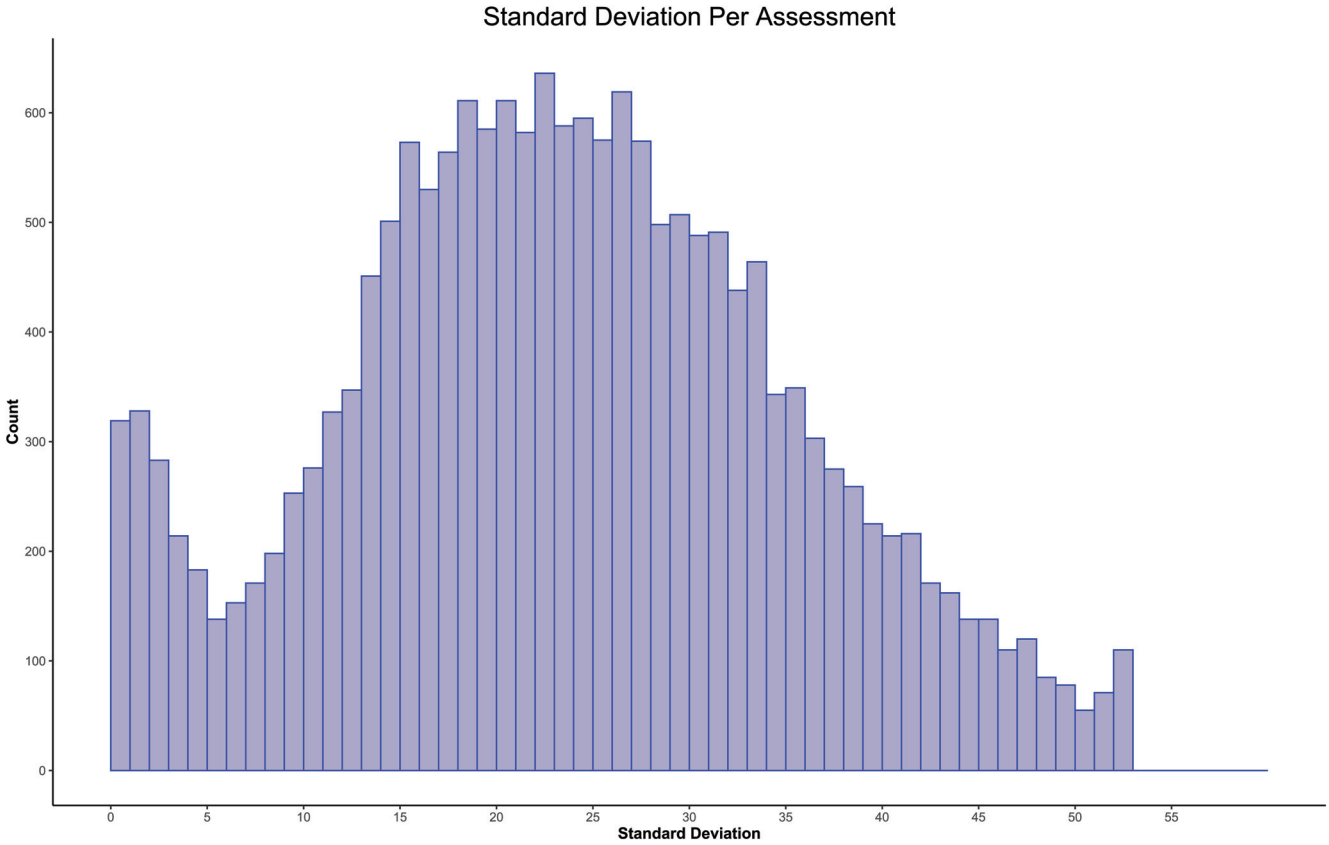
Determining Careless Responses

Time to Complete Each Assessment (See Figure 3)

The total TTC range was 7 s–526,387 s. The skew and kurtosis of the distribution were 70.56 and 5,588; respectively. Because of

Figure 6

The Distribution of the Standard Deviation for All Emotion Items per Assessment



Note. Note the zero inflation (to a within EMA *SD* of approximately 5) and a slight increase in counts at the far-right tail (within EMA *SD* > 50). This distribution includes all assessments (*N* = 18,093). See the online article for the color version of this figure.

the heavy positive TTC skew, we first applied the upper-fence interquartile rule to excise assessments with exceedingly long time to complete. Before applying the upper IQR fence, the TTC mean was M_{TTC} = 161.30 s (*SD* = 5820.2 s). After applying the upper IQR fence, the TTC mean was M_{TTC} = 31.21 s (*SD* = 13.60 s). This TTC mean after applying the upper IQR fence is more realistic given the number of items per cohort varied between 10 and 12. Examining the distribution of TTC across cohort (after z-scoring TTC and applying the upper fence interquartile rule; Figure 8) indicates that the distribution of TTC in our sample was similar regardless of the number of items.

Time to Complete Each Item (See Figure 5)

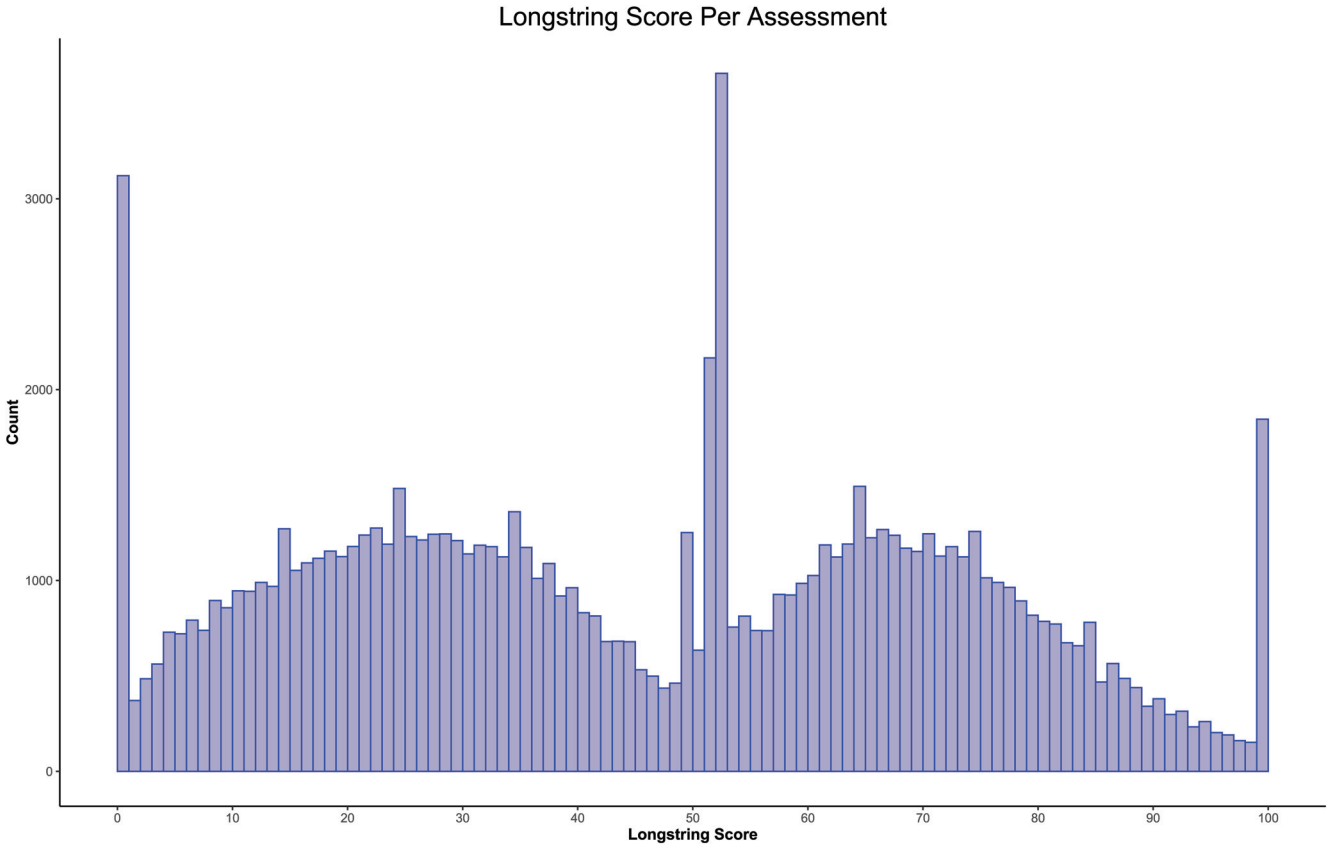
Using Qualtrics, TPI cannot be precisely extracted. Thus, we estimate the TPI = TTC/number of items.

The skew and kurtosis of the TPI distribution were 71.10 and 5,623.76, respectively. As with TTC, because of the heavy positive TPI skew, we first applied the upper interquartile fence to excise assessments that took exceedingly long to complete. Before applying the upper fence, the TPI mean was M_{TPI} = 14.99 s (*SD* = 545.93 s). After applying the upper fence, the TPI mean was M_{TPI} = 2.88 s (*SD* = 1.25 s). TPI and the postquartile-removal mean of TPI are shown in Tables 4B and 4E. Examining the distribution of TPI across cohort (after z-scoring TPI; Figure 9) indicates that the TPI distribution was similar regardless of the number of items contained within each assessment.

We then assessed whether an overly fast TPI was also "careless." First, given the TPI distribution (see Figure 5), we selected TPIs at and below the median TPI (Table 4E). Based on the data, we chose low TPI cutoffs of ≤ 1 s, 1.5 s, 2 s, and 3 s. Using these exclusion criteria, the following number of assessments would be excluded: TPI ≤ 1 s: 372 (2.05%, cumulative: 2.05%) EMAs, 1 s < TPI ≤ 1.5 s: 1,221 (6.75%, cumulative: 8.80%) EMAs, 1.5 s < TPI ≤ 2 s: 2,933 (16.21%, cumulative: 25.02%) EMAs, and 2 < TPI ≤ 3 s: 5,916 (32.70%, cumulative: 57.71%) EMAs.

However, this analysis only tells us the percent of EMAs removed with each cutoff window, it does not inform whether a TPI cutoff was associated with a response that was implausible and perhaps careless. To estimate whether EMAs completed quickly may be careless, we examined the correlation between the psychometric antonyms anxious and relaxed, within the TPIs windows described above. First, the correlation between anxious and relaxed in the entire sample was $r = -.5236$. We then examined the correlation between anxious and relaxed at the four "fast" TPI widows. The correlation between anxious and relaxed were as follows: TPI ≤ 1s: $r = -.0903$; between 1s < TPI ≤ 1.5s: $r = -.3701$; between 1.5 s < TPI ≤ 2 s: $r = -.5822$; and 2 s < TPI ≤ 3 s: $r = -.5783$. As can be seen in Figure 10, the relationship

Figure 7
Longstring Distribution for All Emotion Item Responses per Assessment



Note. See the online article for the color version of this figure.

between relaxed and anxious when an EMA was completed at a rate of less than 1 s per item is virtually flat, indicating that in those responses, participants are not indicating that increases in feelings of being anxious bear any relationship to their feeling of being relaxed. This is unlikely and suggests that the response may be careless. At slightly slower TPIs, for example between 1 s and 1.5 s, the association between anxious and relaxed becomes inverse, albeit shallower than seen in the entire sample. At even slower TPIs, the relationship begins to look much more typical of the entire sample. These results imply that EMAs in which a participant takes less than 1 s to consider their response to each item is likely careless. Some what higher TPIs may also be careless, but it is less certain that they are.

We also explored whether excessively long TPIs (after applying the interquartile rule) were associated with implausible responses. To explore this, we selected 10 s and 20 s per item as an upper fence cutoff. Using the window 10 s $\leq$ TPI $<$ 20s identified 506 (2.80%) EMAs. Selecting TPI $\geq$ 20 s identified 355 (1.96%) EMAs. As above, we examined the correlation between the items anxious and relaxed at these TPIs. The 10 s $\leq$ TPI $<$ 20 s window yielded a correlation of $r = -.3822$; and TPI $\geq$ 20 s yielded a correlation of $r = -.3441$ (see Figure 11). It appears that the association between anxious and relaxed at these higher TPIs are not dissimilar from the entire dataset and thus may not be a useful additional indicator of careless responses once the upper fence is applied.

Within EMA Standard Deviation

In addition to the speed at which an individual completes an EMA, the standard deviation of responses to an individual EMA may also be an important indicator of a careless response (see Figure 6). The skew and kurtosis of the distribution of the *SD* per EMA were .1100 and $-.4064$; respectively. Because there was not a significant skew to the within EMA *SD* distribution and because EMA means and *SD* were similar using the dataset following the application of the upper interquartile fence (Tables 4C and 4F), we did not apply the upper interquartile fence to this analysis.

While the data are relatively normally distributed, there is an initial zero-inflation of within EMA *SD* between 0 and 5. The presence of at least two distributions here suggests that two different generative processes may be occurring, one which generates an EMA with a $SD \leq 5$ and another process which results in an EMA with a $SD > 5$. We confirmed this by calculating the derivative at each *SD* level of the density function (difference between subsequent *SD* levels in the histogram) and identified the location where the slope of the derivative changes from 0 to positive. This

Table 3
Correlations Between All the Possible Pairings of EMA Items

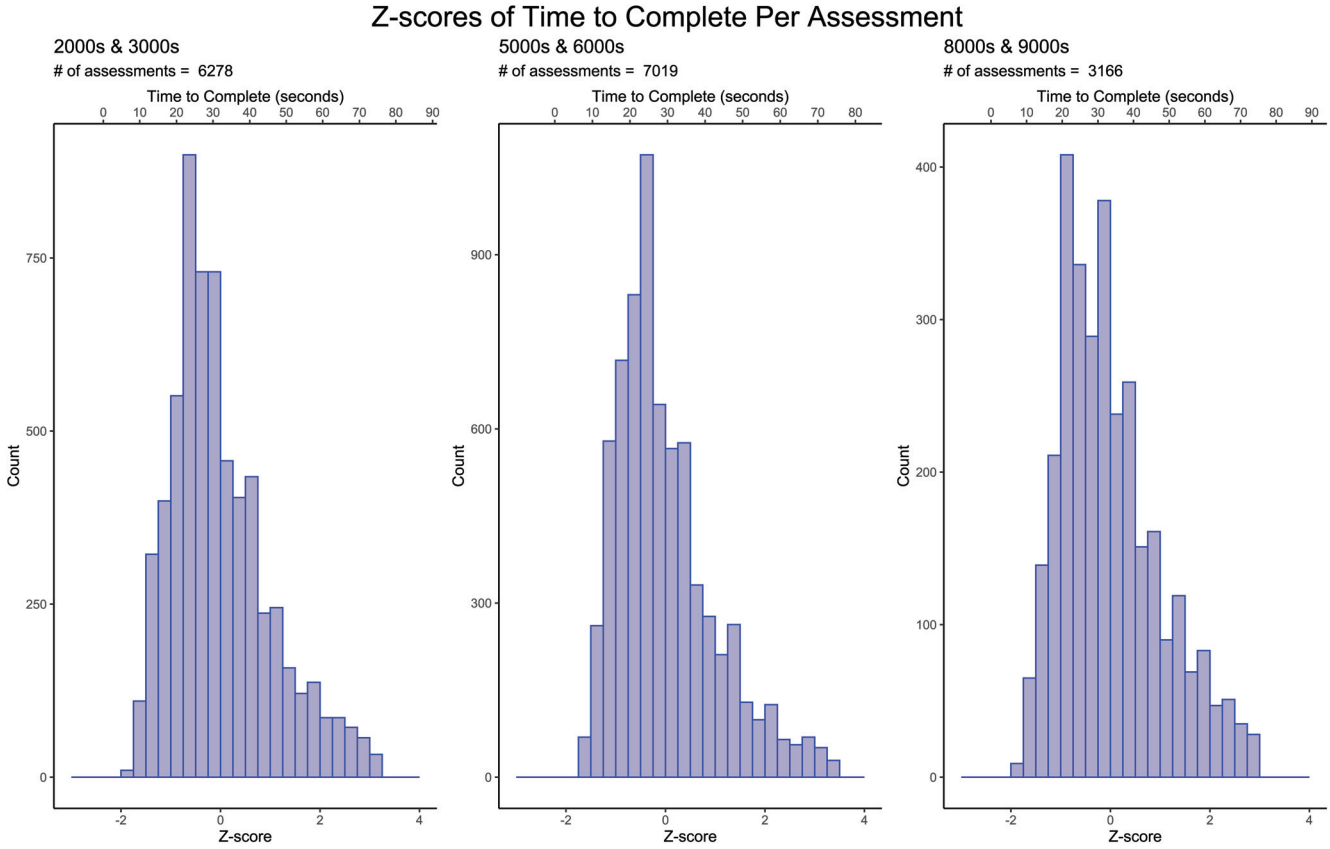
Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Sleep quality	1														
2. Hours of sleep	NA	1													
3. Happy	0.31	0.18	1												
4. Excited	0.21	0.18	0.61	1											
5. Content	0.30	0.16	0.73	0.52	1										
6. Attentive	0.29	0.14	0.36	0.32	0.26	1									
7. Relaxed	0.30	0.17	0.58	0.41	0.58	0.26	1								
8. Sad	NA	−0.11	−0.66	−0.38	−0.51	−0.27	−0.49	1							
9. Tired	NA	−0.31	−0.41	−0.36	−0.35	−0.35	−0.35	0.37	1						
10. Upset	−0.21	−0.11	−0.56	−0.28	−0.53	−0.21	−0.46	0.80	0.36	1					
11. Irritable	−0.21	−0.13	−0.50	−0.28	−0.48	−0.23	−0.45	0.64	0.41	0.69	1				
12. Stressed	−0.17	−0.13	−0.42	−0.29	−0.42	−0.21	−0.53	0.50	0.38	0.52	0.52	1			
13. Anxious	−0.19	−0.12	−0.42	−0.21	−0.41	−0.14	−0.52	0.58	0.37	0.58	0.54	0.63	1		
14. Nervous	−0.17	NA	−0.35	−0.14	−0.37	−0.09	−0.46	NA	NA	0.58	0.53	0.69	0.75	1	
15. Sluggish	−0.32	NA	−0.29	−0.17	−0.25	−0.40	−0.20	NA	NA	0.40	0.44	NA	0.34	0.31	1

Note. EMA = ecological momentary assessment. Because not all cohorts received the same items, some correlations could not be calculated, shown by "NA" (not applicable).

identified an *SD* of 6.16, is near a within EMA *SD* = 5. In addition, there appears to be a slight increase in this distribution around *SD* = 50.

Given the distribution, we hypothesized that a response might be careless if it has an $SD \leq 5$ or if it has an $SD > 50$. A $SD \leq 5$ would indicate a severely restricted range of responses within an EMA, whereas an $SD > 50$ would indicate uniformly polarized responses. We thus performed the same analysis as with TPI, but now at several *SD* windows to determine the number of EMAs that would be identified at each *SD* window: $SD \leq 1$: 319 (1.76%,

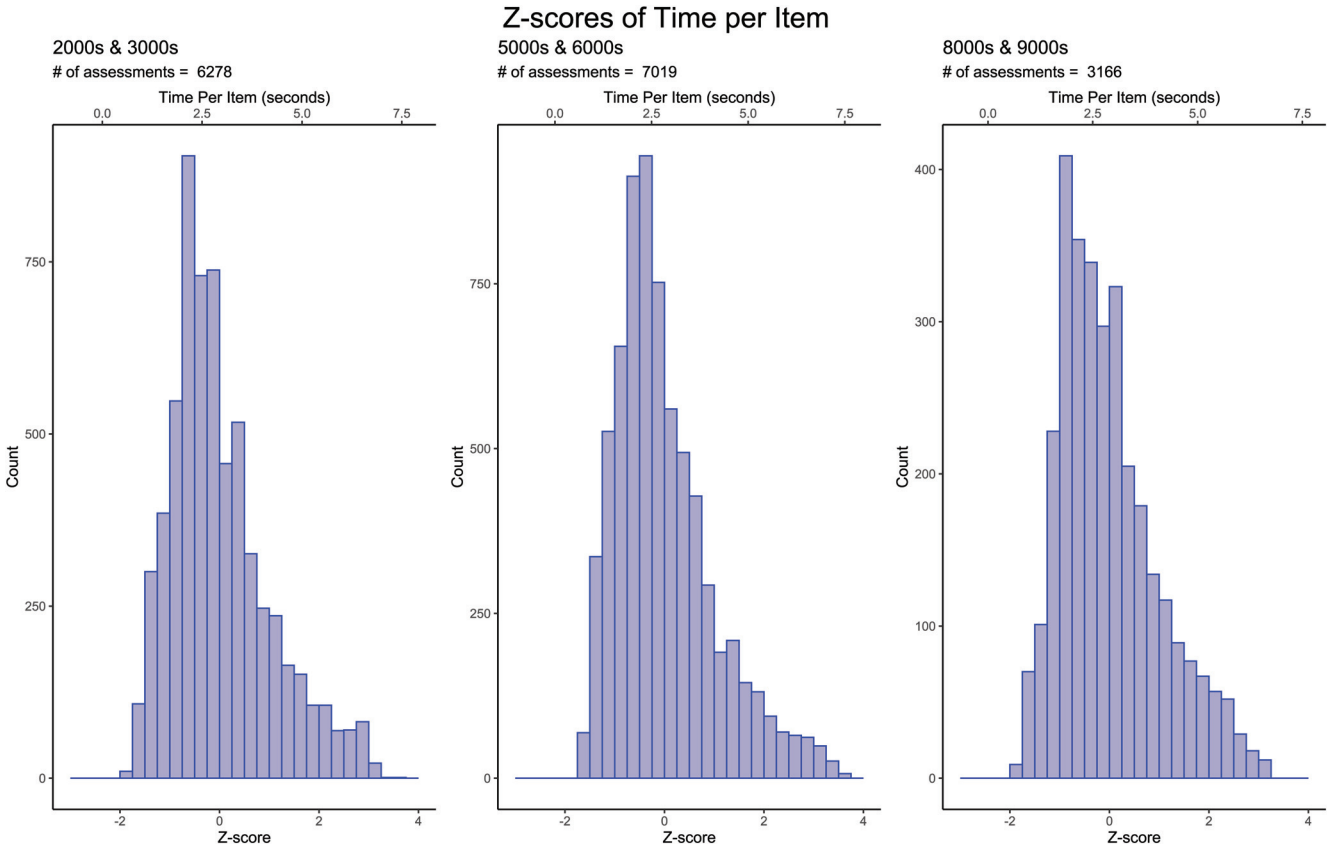
Figure 8
The Distribution of the z-Scored Time to Complete (TTC) per Assessment After the Upper Interquartile Fence Was Applied



Note. Cohorts are grouped such that in each panel, participants completed the identical item set (Table 2). The number of total assessments for each panel is also labeled. X-axis (bottom) is the *z*-scored TTC; x-axis (top) is the raw TTC. See the online article for the color version of this figure.

Figure 9

Distribution of the z-Scored Time per Item (TPI) After the Upper Interquartile Fence Was Applied



Note. Cohorts are grouped such that in each panel, participants completed the identical item set (Table 2). The number of total assessments for each panel is also labeled. X-axis (bottom) is the *z*-scored TPI; x-axis (top) is the raw TPI. See the online article for the color version of this figure.

cumulative: 1.76%) EMAs, $1 < SD \leq 2$: 328 (1.81%, cumulative: 3.58%) EMAs, $2 < SD \leq 5$: 680 (3.76%, cumulative: 7.33%) EMAs, $5 < SD \leq 10$: 913 (5.05%, cumulative: 12.38%) EMAs, $10 < SD \leq 20$: 4,765 (26.34%, cumulative: 38.72%) EMAs, $20 < SD \leq 30$: 5,785 (31.97%, cumulative: 70.69%) EMAs, $30 < SD \leq 50$: 5,067 (28.01%, cumulative: 98.70%) EMAs, $SD > 50$: 236 (1.30%, cumulative: 1.30%) EMAs.

As with the TPI analyses, we followed this initial step by comparing the correlation between ratings of anxious and relaxed in the full sample to those at each *SD* cutoff (see Figure 12). As a reference, with no TPI cutoff across the 18,093 assessments, the correlation between anxious and relaxed was $r = -.5236$. The correlations between anxious and relaxed at the within EMA *SD* windows were: $SD \leq 1$: $r = .9923$ (Figure 12A); $1 < SD \leq 2$: $r = .8940$ (Figure 12B); $2 < SD \leq 5$: $r = .8560$ (Figure 12C); $5 < SD \leq 10$: $r = .4145$ (Figure 12D); $10 < SD \leq 20$: $r = -.1246$ (Figure 12E); $20 < SD \leq 30$: $r = -.3917$ (Figure 12F); $30 < SD \leq 50$: $r = -.6704$ (Figure 12G). Results from these correlations show that the relationship between anxious and relaxed do not begin to resemble the raw data until the *SD* is greater than 20. Furthermore, at low *SD*s, the correlation between anxious and relaxed is highly positive. Using these quantitative values to derive an appropriate cutoff would likely suggest an $SD \leq 20$ as useful for identifying careless responses. However, a qualitative examination of the scatterplots in Figure 12, make it clear that there are several responses that resemble a plausible response such that the response to anxious and relaxed are inversely related. In our data, if we were to rely solely on the quantitative results and define a "good" response as any response with a within EMA $SD < 20$, we would remove 38.71% of all responses. Sole reliance on correlation between anxious and relaxed as the criterion would remove a substantial percent of the data and could impose an artificial restriction on the range of responses by requiring all "good" EMAs to rate anxious and relaxed as maximally inverse to one another. Given the literature that the experiencing of positive and negative emotions simultaneously may be possible (Larsen et al., 2001), moderate within EMA *SD*s may be plausible and may account for the quite high percent of EMAs with an *SD* below 20. Further, given the clear presence of two generative processes in the *SD* distribution around 5, we elected to select the more conservative within EMA $SD \geq 5$ as the suggested cutoff.

We then investigated the relationship between anxious and relaxed at $SD > 50$ due to the small inflation of values visible at the right tail in Figure 6. Using a cutoff of $SD > 50$, 236 assessments were identified (1.3% of the assessments). The correlation between anxious and relaxed at this cutoff was very high and

Table 4
Descriptive Statistics of Time to Complete (TTC) EMA, Time per Item (TPI), and Within EMA SD

A	Time to complete per assessment (seconds)		
Groups	2,000 s and 3,000 s	5,000 s and 6,000 s	8,000 s and 9,000 s
Number of Items	11	10	12
Mean (*SD*)	269.6 (7,785)	112.4 (4,992)	51.39 (158.5)
Median	31	28	33
B	Time per item per assessment (seconds)		
Groups	2,000 s and 3,000 s	5,000 s and 6,000 s	8,000 s and 9,000 s
Number of Items	11	10	12
Mean (*SD*)	24.54 (707.7)	11.05 (499.2)	4.45 (14.13)
Median	2.81	2.7	2.82
C	All item standard deviation per assessment		
Groups	2,000 s and 3,000 s	5,000 s and 6,000 s	8,000 s and 9,000 s
Number of Items	11	10	12
Mean (*SD*)	23.18 (11.40)	23.37 (12.09)	25.33 (11.25)
Median	23.04	23.06	24.71
D	Time to complete per assessment (seconds) postquartile removal		
Groups	2,000 s and 3,000 s	5,000 s and 6,000 s	8,000 s and 9,000 s
Number of Items	11	10	12
Mean (*SD*)	32.16 (13.76)	29.56 (13.39)	33.62 (14.14)
Median	29	26	31
E	Time per item per assessment (seconds) postquartile removal		
Groups	2,000 s and 3,000 s	5,000 s and 6,000 s	8,000 s and 9,000 s
Number of Items	11	10	12
Mean (*SD*)	2.94 (1.26)	2.84 (1.30)	2.91 (1.20)
Median	2.64	2.55	2.67
F	All item standard deviation per assessment postquartile removal		
Groups	2,000 s and 3,000 s	5,000 s and 6,000 s	8,000 s and 9,000 s
Number of Items	11	10	12
Mean (*SD*)	23.21 (11.59)	23.24 (12.20)	25.19 (11.41)
Median	23.10	22.90	24.46

Note. EMA = ecological momentary assessment. Subtables are split into the cohort groups (Table 2) and display the number of EMA items, *M* (*SD*), and median for each descriptive statistic. Rows A-C are descriptive statistics including all data, prior to quartile removal. Rows D-F include only those EMAs surviving quartile removal.

negative ($r = -.8721$; Figure 12H). This correlation reflects the directionality of the raw data, but the effect size is stronger than that observed in the full data because an EMA with an $SD > 50$ has several "extreme" values (rated either 0 or 100) and the leverage these extreme values have a strong effect on the correlation between anxious and relaxed. Whether assessments with such high within EMA *SD* may be careless is worth further exploration, however, the longstring approach (described below) may capture this process as well.

Within EMA Mode

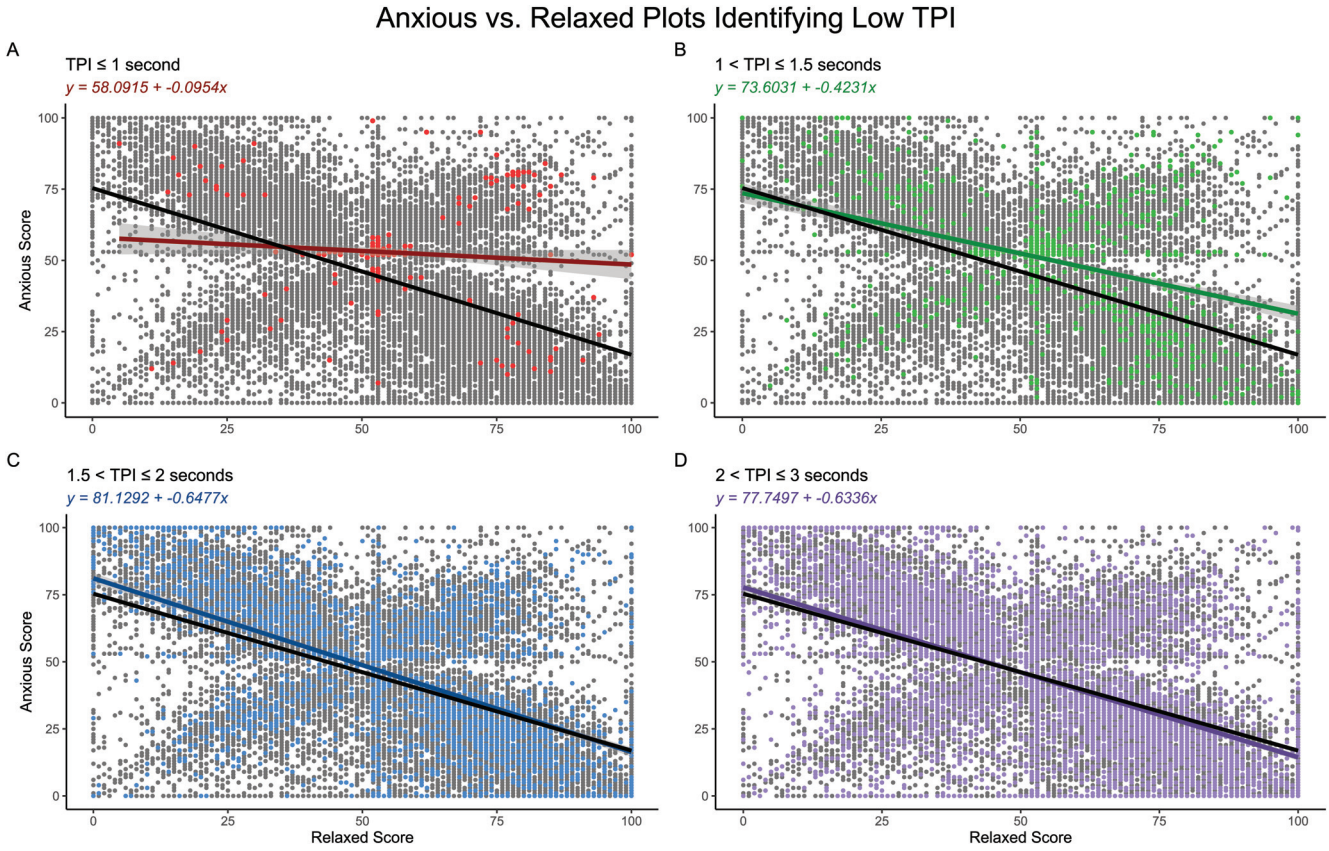
The skew and kurtosis of the mode of each EMA ("longstring" value; Figure 7), was .0159 and −1.015, respectively. The distribution of the longstring response per assessment was bimodal, with inflation around the item scores 0, 50–54, and 100. This indicates that there is a tendency for participants to endorse many items as not present (0), very strong (100), or in the middle (50–54). Careless responses may be observable if the percent of the within EMA items at the mode is greater than half of the number of assessment items. This may be particularly useful if the EMA is assessing multiple constructs, particularly if constructs are theorized to be inversely related (e.g., positive and negative affect), because it would indicate that more than half of the items were endorsed at identical level.

We assessed whether the percent of items at the mode could be a third metric to flag a response as "careless." Because the EMAs used here assessed two constructs (positive and negative affect), we hypothesized that if the mode of an EMA was selected more than 50%, 60% or 70% of the time, that EMA may be identifiable as careless. Using these exclusion criteria, the following number of assessments would be excluded: % items at mode ≥ 50%: 1,772 (9.79%) items at mode ≥ 60%: 1,129 (6.24%), % items at mode ≥ 70%: 696 (3.85%).

However, this analysis only tells us the percent of EMAs removed with each cutoff, it does not inform whether a % of items at the mode cutoff was associated with a response that was

Figure 10

Comparing the Relationship Between Psychometric Antonyms at Fast Time per Item (TPI) Windows



Note. The plots in this figure compare the correlation between the response to "anxious" and the response to "relaxed," broken down into different fast TPI windows (colored lines). Each plot also contains the overall regression line including all items (solid black line; y = 75.3696–0.5848x). (A) Red points illustrate where TPI ≤ 1s. A total of 372 assessments were identified (2.06%). (B) Green points illustrate where TPI > 1s and TPI ≤ 1.5s. A total of 1,221 assessments were identified (6.75%). (C) Light blue points illustrate where TPI > 1.5 s and TPI ≤ 2 s. A total of 2,933 assessments were identified (16.21%). (D) Purple points illustrate where TPI > 2 s and TPI ≤ 3 s. A total of 5,916 assessments were identified (32.70%). As can be seen, a TPI ≤ 1 s typically results in an implausible response. At the other windows plausible responses are substantially more common. See the online article for the color version of this figure.

implausible. We examined the correlation between the psychometric antonyms anxious and relaxed, within the % items at mode described above. The correlation between anxious and relaxed were as follows: % items at mode ≥ 50%: $r = -.5054$ (Figure 13A); % items at mode ≥ 60%: $r = -.0463$ (Figure 13B); % items at mode ≥ 70%: $r = .2630$ (Figure 13C; all $p < .001$). As can be seen in Figure 13B, the relationship between relaxed and anxious is virtually flat when the percent of items at the mode is ≥ 60%. This indicates that in these responses, participants are not indicating that increases in feelings of being anxious bear any relationship to their feelings of being relaxed. This is unlikely and suggests that the response may be careless.

Integrating Metrics to Determine Careless Responses

It may be that combining three metrics: TPI, *SD*, and % items at the mode provides the best and most comprehensive way to determine careless responses. We use the logical Boolean "OR" and "AND," and we describe results below.

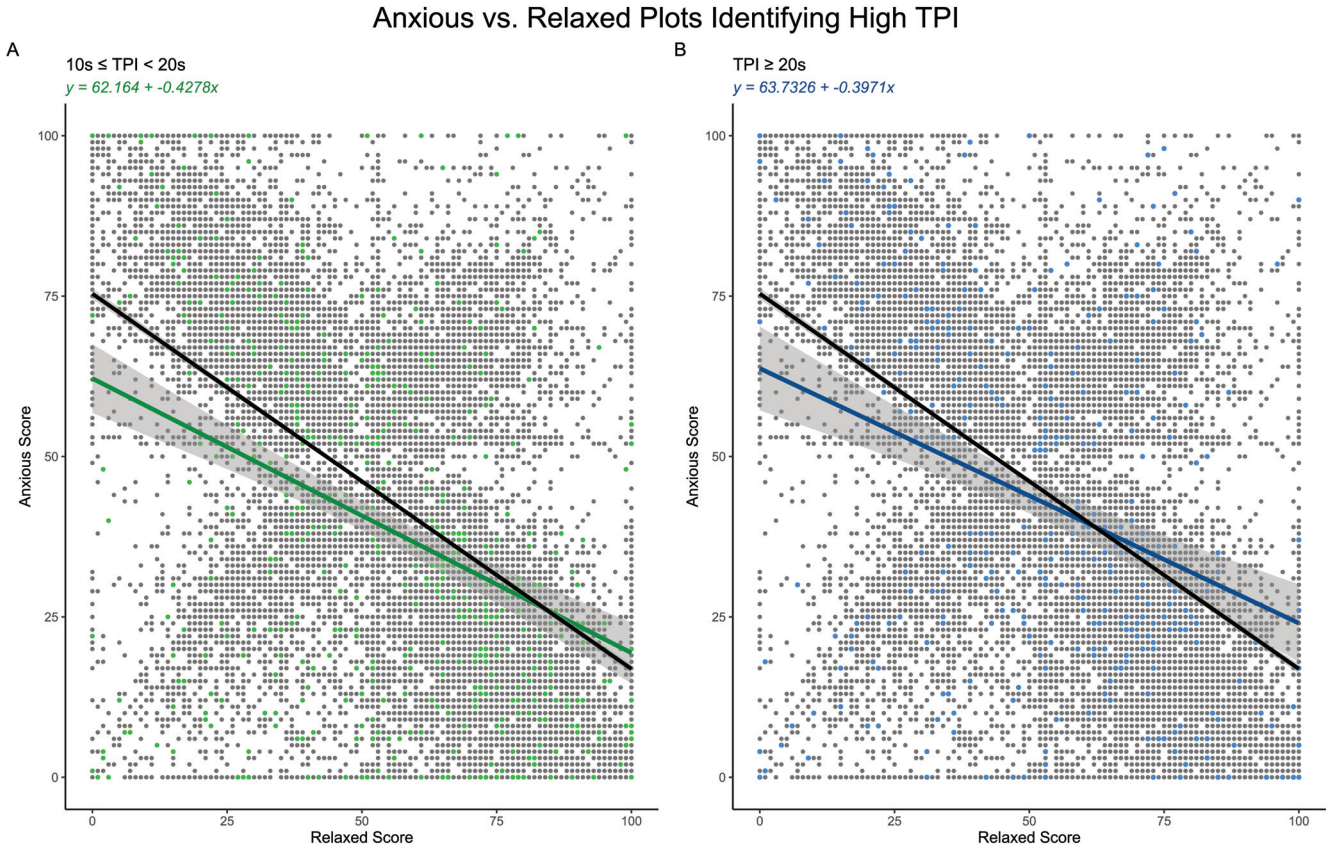
Boolean OR

A data-driven approach using the correlation between responses to anxious and relaxed yielded a TPI cutoff of ≤ 1 s, an *SD* ≤ 20, and a % items at mode ≥ 60% as three criteria. Flagging all assessments satisfying any of these as careless would remove 7,171 (39.63%) EMAs (Figure 14A). Applying this approach would likely identify all careless responses but would be highly restrictive and could also identify several responses that are not careless. It is also quite clear that there are a many responses falling along the top-left to bottom-right diagonal, which would otherwise be considered plausible responses.

Second, we applied the same Boolean "OR" logic but used the more conservative *SD* cutoff: a within EMA *SD* ≤ 5 instead of *SD* ≤ 20. This captures the initial inflation/hurdle observed in Figure 6 and we theorize more directly captures a distinctive generative when completing an EMA. Applying this approach (along with the TPI ≤ 1 s and % items at mode ≥ 60%) would remove of 1,694 (9.36%)

Figure 11

Comparing the Relationship Between Psychometric Antonyms at Slow Time per Item (TPI) Windows



Note. The plots in this figure compare the correlation between the response to "anxious" and the response to "relaxed," broken down into two different slow TPI windows (colored lines). Each plot also contains the overall regression line including all items (solid black line; y = 75.3696–0.5848x). (A) Green points illustrate where 10 ≤ TPI and < 20 s. A total of 506 out of 18,093 assessments are identified (2.80%). (B) Blue points illustrate where the TPI > 20 s per item. A total of 355 out of 18,093 assessments are identified (1.96%). See the online article for the color version of this figure.

EMAs in our data (Figure 14B). While less severe of a penalty, this approach clearly identifies poor responses and is less likely to remove potentially good data.

We also explored whether these flagging metrics were redundant. Using conjunction (Boolean AND) for TPI ≤ 1s, *SD* ≤ 20, % items at mode ≥ 60%, identified 218 responses (1.20%; Figure 14C). Using this same conjunction for TPI ≤ 1 s, *SD* ≤ 5, % items at mode ≥60%, identified 200 responses (1.11%; Figure 14D). This indicates that these flagging metrics identify unique sources of possible carelessness, leading us to recommend using Boolean OR when searching for careless responses.

Considering Increases in EMA Completion Efficiency

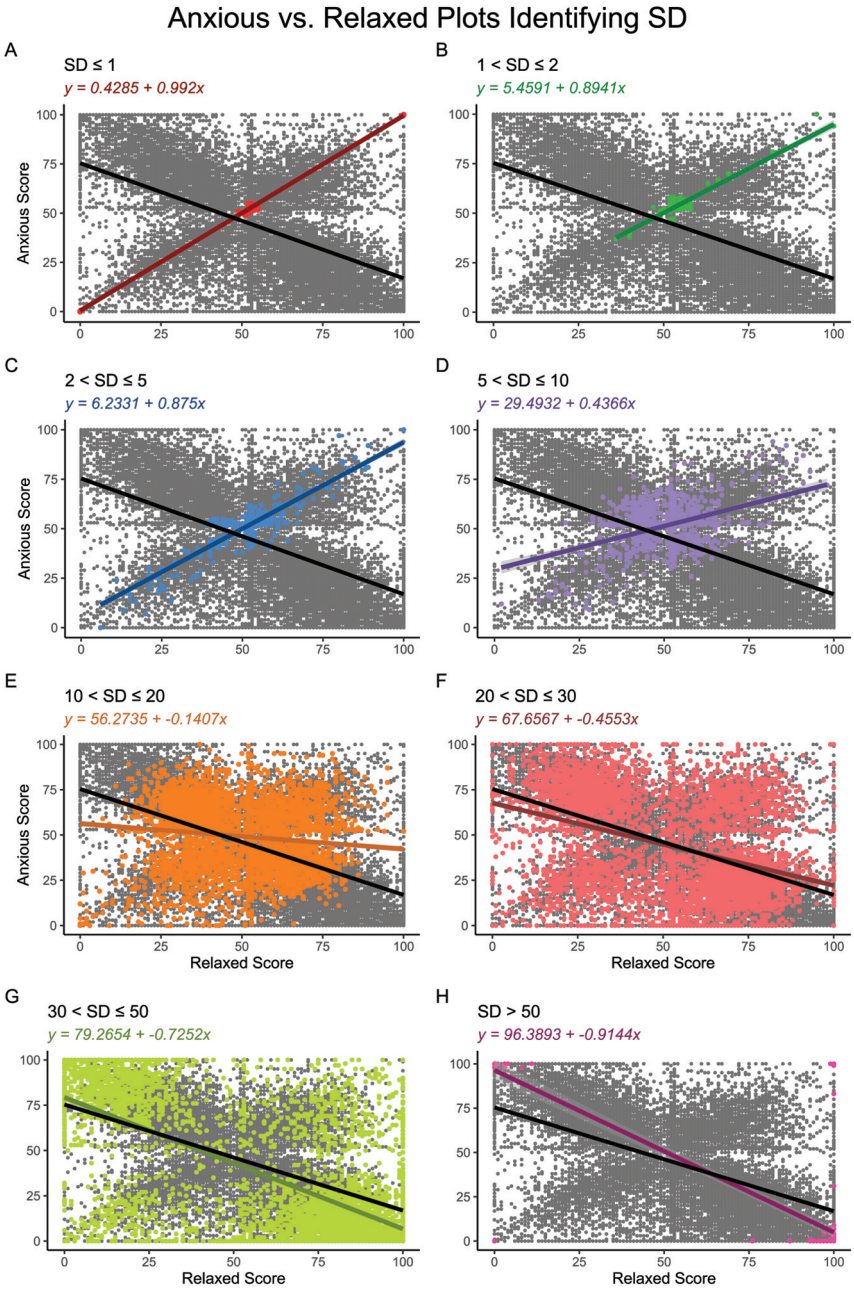
For the majority of EMA studies, identical assessments are sent repeatedly. Thus, even if item-order is randomized, participants may become faster at responding to each EMA. For instance, it was recently demonstrated that individuals respond in a more extreme ("elevated") manner on the first few EMAs, but that this effect attenuates across repeated assessments (Shrout et al., 2018). This complicates implementation of rigid cutoffs for a careless response, and further suggests that using slightly more liberal cutoffs could be warranted as participants become accustomed to completing assessments. To test this, we assessed whether the median TPI of the first five assessments and median TPI of the last five assessments differed (see Figure 15). The median TPI, as opposed to mean TPI was used to reduce the effect of outliers on the data. In this study, participants completed EMAs over a semester (∼4 months), and on average completed 60 EMAs. A visual inspection shows that participants became substantially faster responders across the study period. A paired samples *t*-test corroborated this, indicating that the median TPI for the first five assessments (M = 4.06 s, SD = 1.64) was significantly slower, $t(287) = 14.935$, $p < .001$, than the median TPI for the last five assessments (M = 2.67, SD = 1.12). This suggests that a slightly more conservative cutoff is warranted so as to account for changes in completion efficiency over the study's duration.

Determining Careless Responders

Thus far, our examination has been limited to identifying careless responses, without consideration for how such responses cluster within individuals. Thus, we examined the combination of

Figure 12

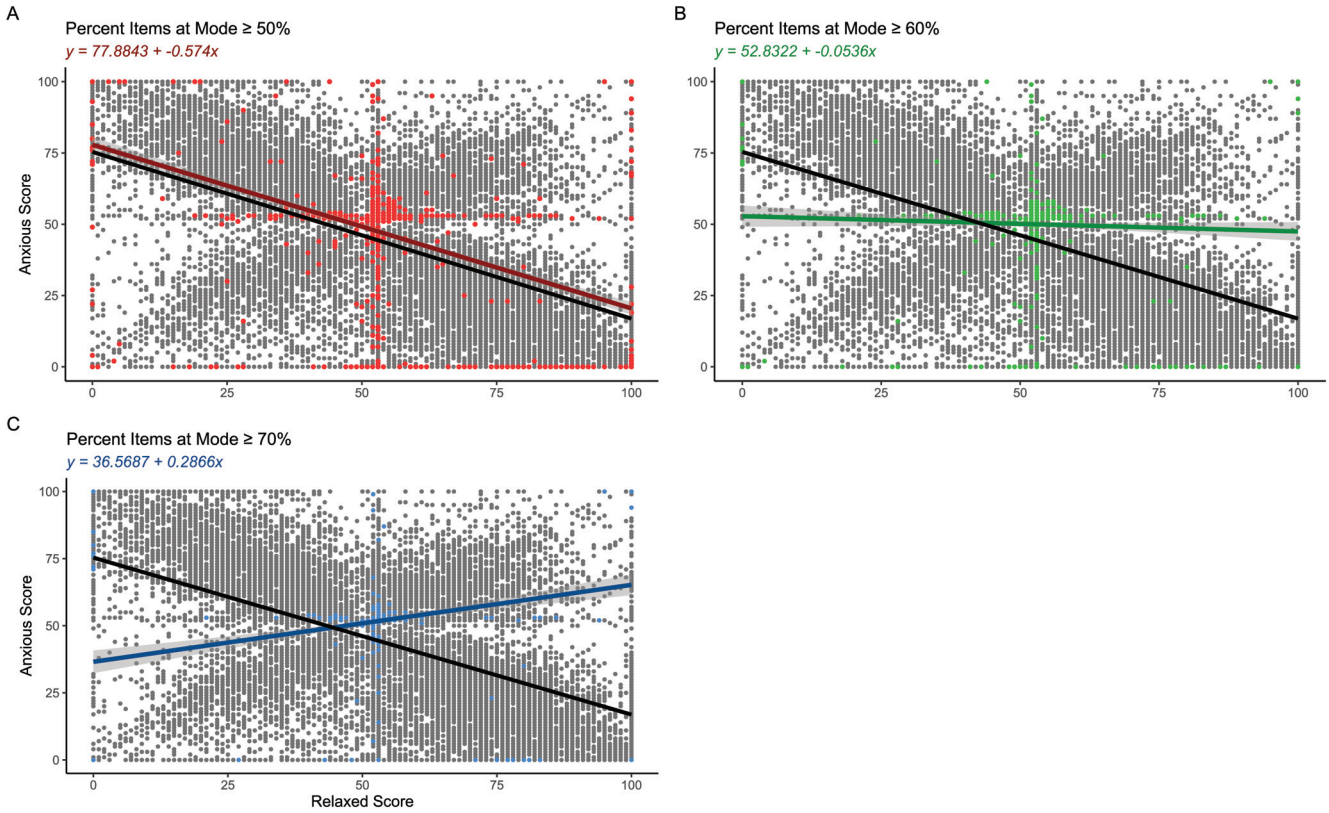
Comparing the Relationship Between Psychometric Antonyms ("Anxious" and "Relaxed") at Different Levels of Within EMA SD



Note. EMA = ecological momentary assessment. Each plot also contains the overall regression line including all items (solid black line; y = 75.3696–0.5848x). (A) Red points indicate where $SD \leq 1$. A total of 319 out of 18,093 assessments are identified (1.76%). (B) Green points indicate where $SD > 1$ and $SD \leq 2$. A total of 328 out of 18,093 assessments are identified (1.81%). (C) Light blue points indicate where $SD > 2$ and $SD \leq 5$. A total of 680 out of 18,093 assessments are identified (3.76%). (D) Light purple points indicate where $SD\ 5 >$ and $SD \leq 10$. A total of 913 out of 18,093 assessments are identified (5.05%). (E) Orange points indicate where $10 <$ and $SD \leq 20$. A total of 4,765 out of 18,093 assessments are identified (26.34%). (F) Coral red points indicate where $SD > 20$ and $SD \leq 30$. A total of 5,785 out of 18,093 assessments are identified (31.97%). (G) Lime green points indicate where $SD > 30$ and $SD \leq 50$. A total of 5,067 out of 18,093 assessments are identified (28.01%). (H) Magenta points indicate where $SD > 50$. A total of 236 out of 18,093 assessments are identified (1.30%). See the online article for the color version of this figure.

Figure 13

Comparing the Relationship Between Psychometric Antonyms At Windows of the Proportion of Items at Mode ("Longstring")



Note. The plots in this figure compare the correlation between the response to "anxious" and the response to "relaxed," broken down into three different cutoffs (colored lines). Each plot also contains the overall regression line including all items (solid black line; y = 75.3696–0.5848x). (A) Red points illustrate where the percent of items at the mode ≥ 50%. A total of 1,772 out of 18,093 assessments are identified (9.79%). (B) Green points illustrate where the percent of items at the mode ≥ 60%. A total of 1,129 out of 18,093 assessments are identified (6.24%). (C) Blue points illustrate where the percent of items at the mode ≥ 70%. A total of 696 out of 18,093 assessments are identified (3.85%). See the online article for the color version of this figure.

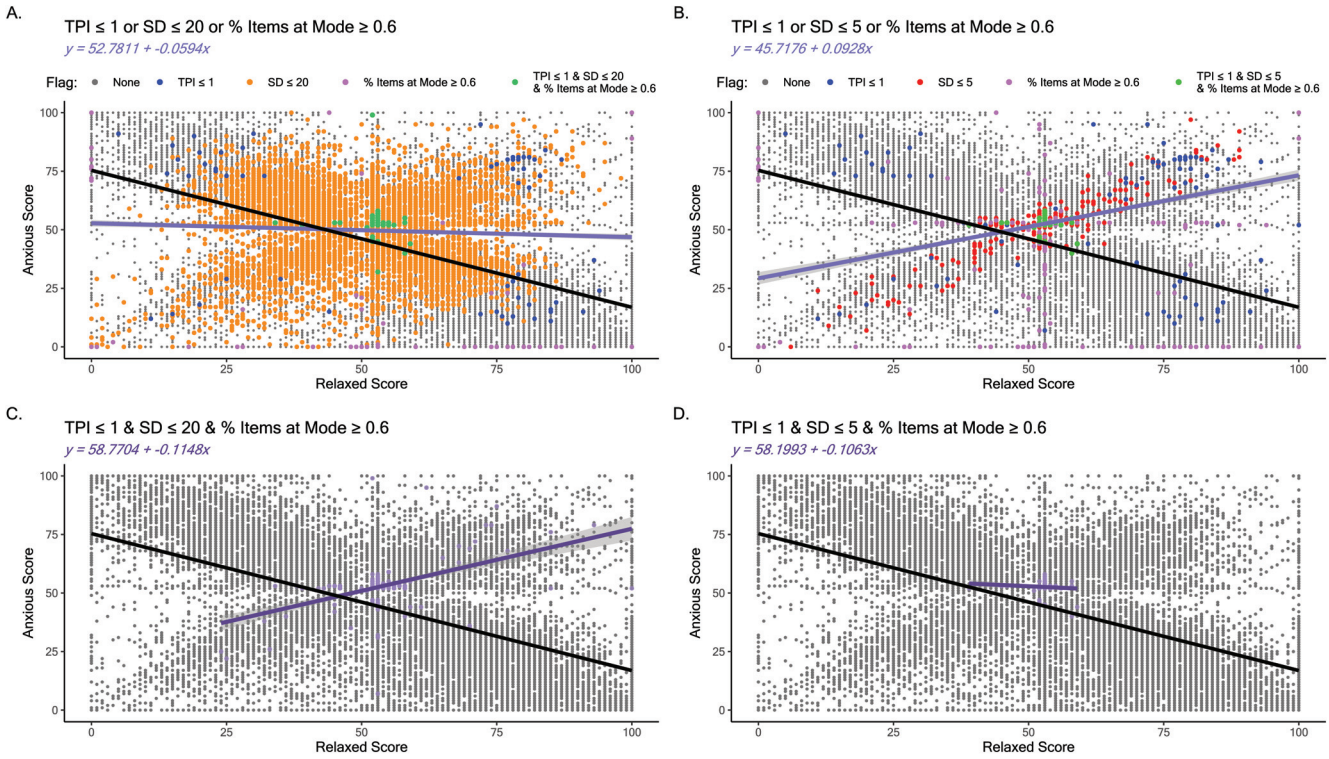
careless response metrics (TPI ≤ 1 s OR within EMA *SD* ≤ 5 OR % items at mode ≥ 60%) nested within participants. Of the 293 participants, 174 (59.39%) participants had at least one EMA response identified as careless by the above criteria. 16 (5.46%) of the 293 participants had 50% of their EMAs flagged as careless and three (1.02%) of the 293 participants had 75% of their EMAs flagged as careless (see Figure 16). Thus, these were the participants who were most consistently providing careless responses during the study. Depending on the leniency of the inclusion criteria, it may be worth considering removing them from group analyses.

We similarly examined the combination of careless response metrics at the more liberal cutoff (TPI ≤ 1 s OR within EMA *SD* ≤ 20 OR % items at mode ≥ 60%) nested within participants. Using this metric, 113 (38.57%) of all 293 participants had 50% of their EMAs flagged as careless and 39 (13.31%) of all 293 participants had 75% of their EMAs flagged as careless (see Figure 17). Thus, these were the participants who were most consistently providing careless responses during the study, albeit at the more liberal level. Determining whether to remove such participants from group analyses is warranted.

In determining whether to remove a participant from analysis it is worth considering how many samples are necessary to have a stable estimate of one's variable of interest. In the case of measures of affective functioning assessed via EMA, this question could be rephrased as, "how many EMAs are necessary to have a stable estimate of mean or standard deviation of positive/negative affect." We examine both mean and standard deviation of positive affect as these affective dynamics measures are frequently explored in the EMA literature (Dejonckheere et al., 2019). For each subject, we randomly sampled a subset of their EMA data (1, 2, . . . , 35 EMAs) and calculated the mean/*SD* of positive affect (PA) for that random sample. We also calculated the mean/*SD* of all EMAs for each subject as their "ground truth" estimate of mean PA and PA *SD*. We then calculated the between subject's rank-order correlation between the ground truth and the random sample. As can be seen in Figure 18, mean PA stabilizes around 15 EMAs, PA *SD* stabilizes around 20–25 EMAs. Thus, when considering whether to remove a participant completely, these criteria (or general approach) may be useful to guide researchers to determine how many EMAs are needed to have a stable trait-like estimate of affective state.

Figure 14

Differential Effects of Combining Metrics (TPI, SD, % Items at Mode) Using Boolean Logic (AND vs. OR) to Flag Careless EMAs on the Correlation Between "Anxious" and "Relaxed"



Note. EMA = ecological momentary assessment. Only assessments with responses for both "anxious" and "relaxed" items were used (N = 18,093). The black line is the linear regression of all data ($r = -0.5236$). (A) Assessments in which TPI ≤ 1 s OR the within assessment $SD \leq 20$ OR % of items at the mode ≥ 60%. These flagged assessments are highlighted in different colors based on the combination of criteria. A total of 7,171 (39.63%) assessments are identified by a combination of either criteria. The purple line is the linear regression of only those points identified ($r = -0.0531$). (B) Assessments in which TPI ≤ 1 s OR the within assessment $SD \leq 5$ OR % of items at the mode ≥ 60%. The assessments are highlighted in different colors based on which combination of imposed criteria they meet. A total of 1,694 (9.36%) assessments are identified by a combination of either criteria. The purple line is the linear regression of only those points identified ($r = 0.0830$). (C) Assessments in which TPI is ≤ 1 s AND the within assessment $SD \leq 20$ AND % of items at the mode ≥ 60%. These assessments are highlighted as purple dots. A total of 218 (1.20%) assessments satisfy both criteria simultaneously. The dark purple line is the linear regression of only those points identified ($r = -0.0652$). (D) Assessments in which TPI ≤ 1 s AND the within assessment $SD \leq 5$ AND % of items at the mode ≥ 60%. These assessments are highlighted as purple dots. A total of 200 (1.11%) assessments satisfy both criteria simultaneously. The dark purple line is the linear regression of only those points identified ($r = -0.1140$). See the online article for the color version of this figure.

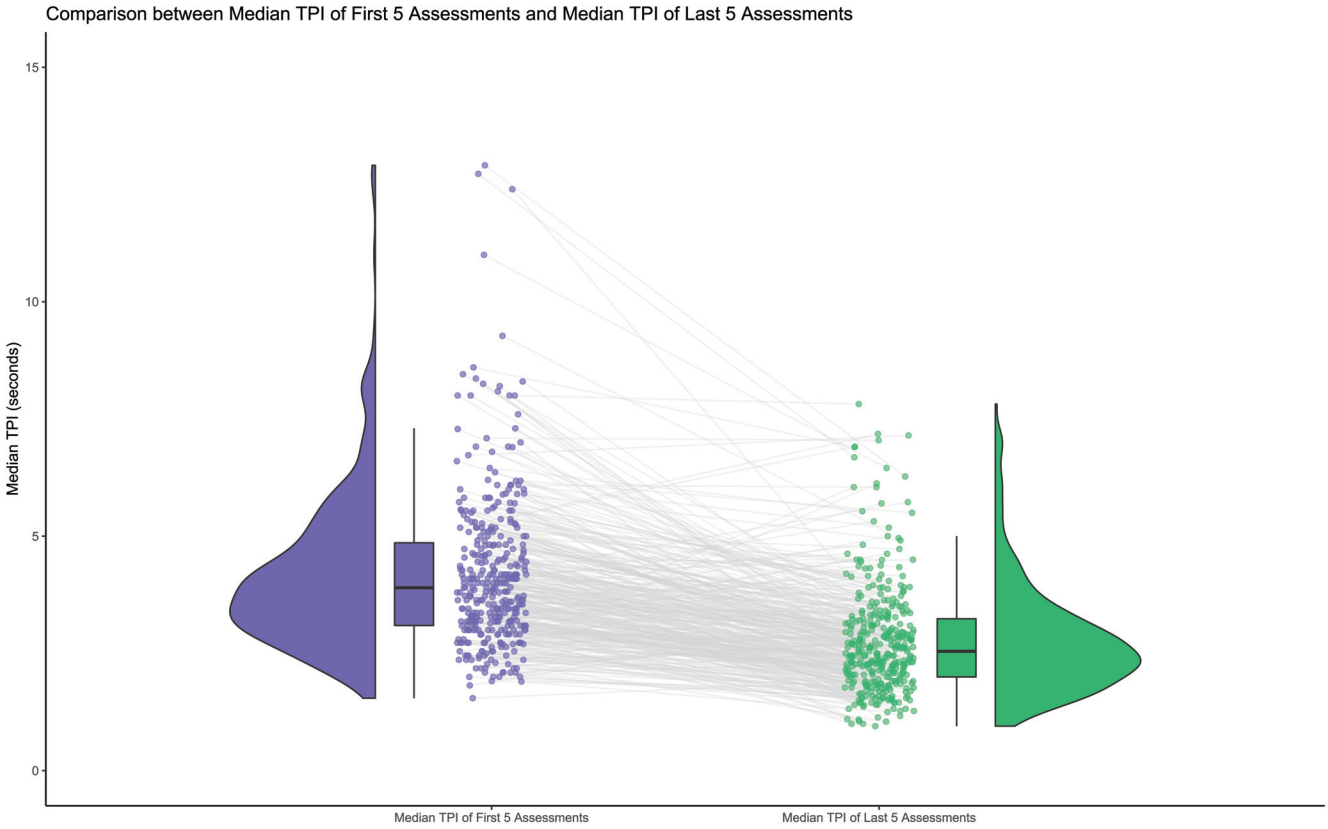
Discussion

Using EMA to advance affective and clinical science has blossomed in the last decade. As this research literature has matured it is imperative to maximize data quality. A majority of EMA studies rely exclusively on compliance to determine whether individual subjects will be included in analyses. However, a long-standing research history, from self-report surveys (Johnson, 2005; McCabe et al., 2012; Meade & Craig, 2012), to behavioral data via RT (Christensen et al., 2003), indicates that determining the quality of individual responses are critical to ensuring the reliability and validity of results. Taking advantage of a large corpus of EMA responses, we determined thresholds for identifying careless responses and whether certain individuals displayed frequent careless response patterns—a so-called, careless responder—and should be considered for removal from subsequent group analyses.

Our results demonstrate the utility of criteria for determining whether individual EMA responses should be included or considered careless and excluded from analyses. Three metrics were particularly useful for identifying careless responses. First, the time taken to complete an EMA item is useful to identify careless responses. For our relatively simple, single-word emotion adjectives (most of which were taken from the PANAS-X), we observed that a time per emotion item of approximately 1 s or less was a reasonable cutoff for a careless response. Broadly, the majority of EMA research uses similar items, assessing similar constructs. However, some EMA studies may assess more abstract concepts or use more complex language. EMAs with more complicated assessments may require longer TPI/TTC cutoffs to be considered valid, which is why we emphasize the importance of data-driven approaches (including data visualization) for determining the cutoff during data-cleaning.

Figure 15

Comparing the Median TPI of the First Five Assessments (Purple, on Left) to the Median TPI of the Last Five Assessments (Green, on Right) for Every Participant



Note. TPI = time per item. The half violin plot shows the distribution of data points in order to better visualize the clustered data. A boxplot of the data is also included. The gray lines represent individual participants change in TPI. See the online article for the color version of this figure.

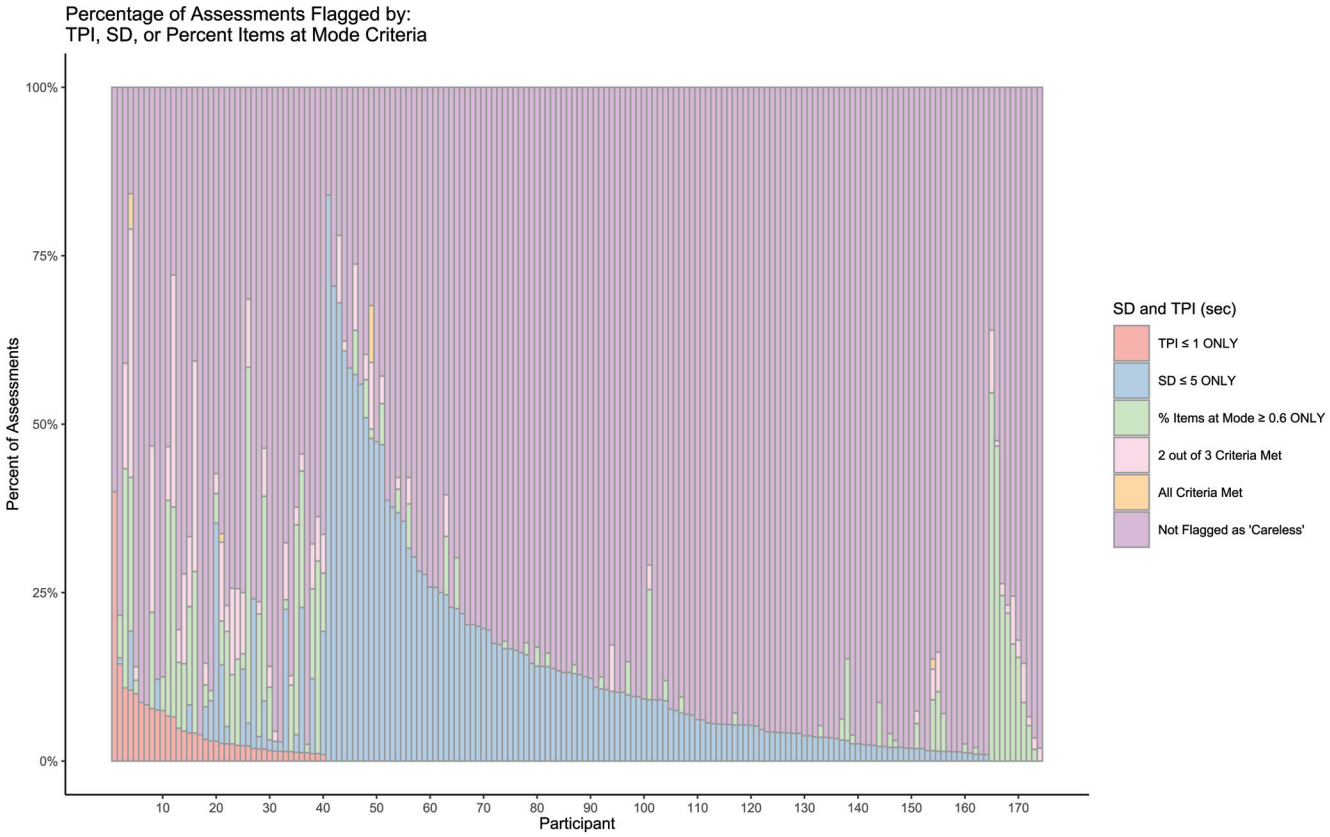
Second, measures of within assessment standard deviation across items was highly useful in determining a possible careless response. In our data, the double-peaked distribution of the within EMA *SD* with one peak close to an *SD* equal to zero suggested two independent generative processes. These two processes appeared to be separated around a within EMA *SD* equal to 5 which led to our suggestion of a cutoff around 5. While this cutoff was determined qualitatively via visual inspection, we also introduce a method (using the derivative of the density distributions) for identifying where the two distributions separate. Nonetheless, similar to the TPI cutoff, the appropriate within EMA *SD* cutoff may vary across studies based upon the number of constructs being assessed and their theoretical relation with one another. Overall, the within assessment standard deviation is very useful to determine a lower-bound for the amount of variance required for a valid response.

The proportion of items at the mode was our third metric to independently identify careless responses. Our EMAs assessed positive and negative affect, constructs which should vary inversely with one another. Using three difference cutoffs, we found that an EMA containing greater than 60% of the items at the mode was likely careless. At this level, the correlation between ratings of anxious and relaxed was around zero. One particular strength of this metrics was that, while the within EMA *SD* criteria was excellent at identifying careless responses focused at the middle of the response range (∼50 from a possible score of 0–100), the proportion of items at the mode metric identified careless responses much more evenly throughout the response space. Because EMA data are often highly skewed, with responses bunched at the extremes, applying this metric is critical to supplement the within EMA *SD*.

We recommend using all criteria simultaneously to determine a careless response. Using a Boolean logical OR test, we flagged all EMAs satisfying a within EMA $SD \leq 5$, a TPI $\leq$ 1 s, or a percent of items at mode $\geq 60\%$ as careless. Critically, the overlap among these metrics was minimal, indicating that they capture distinct indices of carelessness. Applying these relatively liberal criteria reduced the number of very likely "careless" EMAs while retaining as much of the data as possible. In that vein, given our analysis indicating that participants tend to become more efficient at completing EMAs across time, maintaining a somewhat more liberal cutoff for TPI or other metrics may be more appropriate in lieu of either individualized or time-varying cutoffs, which are far more challenging to implement accurately. Similarly, if modeling

Figure 16

Breakdown of the 121 Participants Who Had at Least One Response That Met Criteria for a Within EMA SD ≤ 5 OR TPI ≤ 1 s OR % of Items at the Mode ≥ 60%



Note. TPI = time per item; EMA = ecological momentary assessment. Each vertical line represents one of the 174 participants, with the percentage of EMA that met the TPI ≤ 1 s only (red), the *SD* ≤ 5 criteria only (blue), the % items at mode ≥ 60% (green), two out of three criteria were met (pink), all three criteria were met (orange), and assessment that satisfy none of the careless criteria (purple). Note that most participants did not meet criteria for a careless response for more than 50% of the responses. See the online article for the color version of this figure.

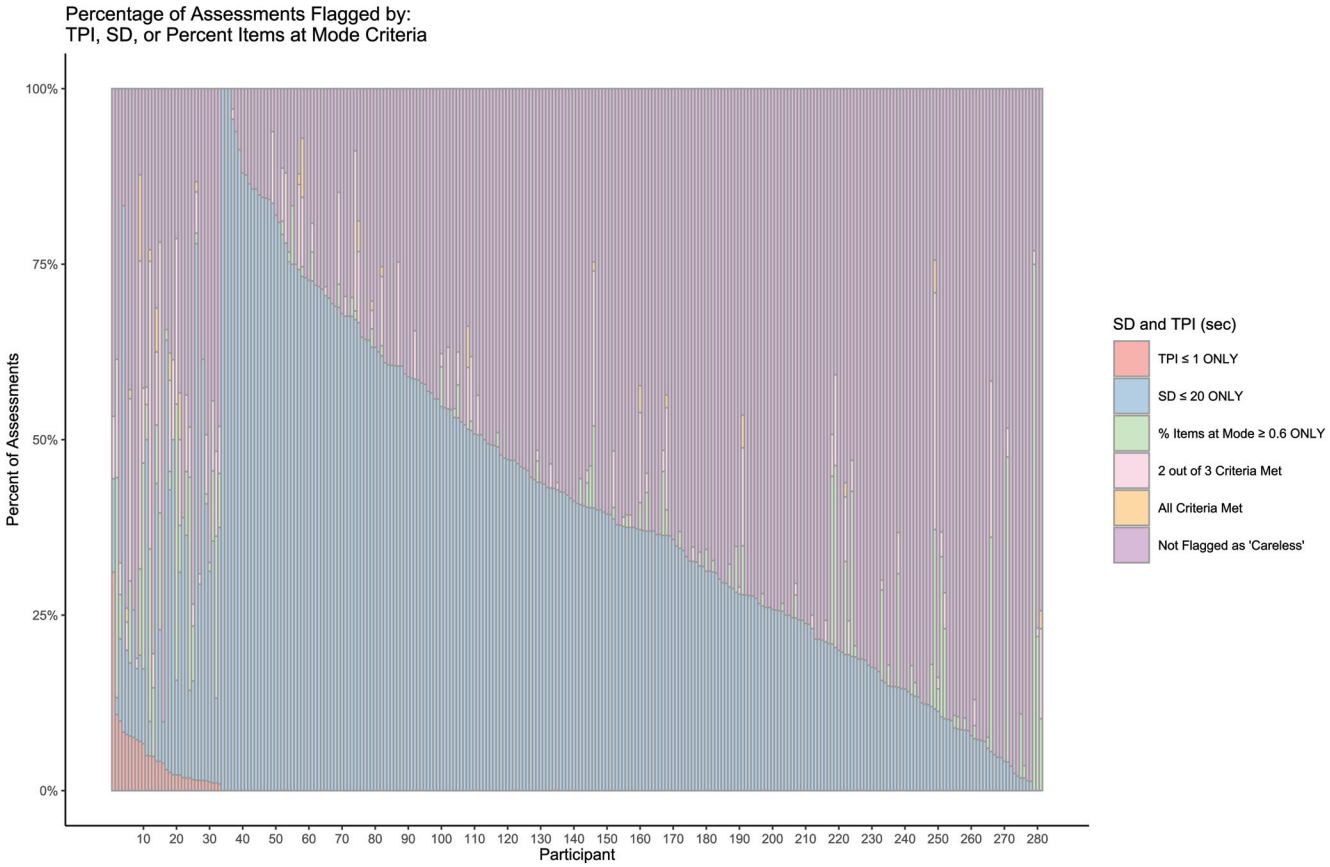
data using hierarchical, or multilevel models, a within-participant predictor for time could be used to remove variance associated with practice effects.

This leads to the question of how to set criteria for identifying careless or noncompliant participants and perhaps be removed from analyses completely. The survey and behavioral (e.g., RT) literatures provide no clear guidelines. The situation is no different in the EMA literature where researchers use compliance rates (typically between 60% and 80%) as a cutoff for inclusion. While these percentages are arbitrary, the rationale is that if a substantial proportion of trials or EMA responses are excluded, the participant either did not understand the task and/or were not paying attention to the study (Congdon et al., 2012). Here, we provide a more quantitative analysis to guide researchers on what might be the minimal number of EMAs for a stable estimate of mean and variability of affect. By randomly sampling subsets of each participant's EMA data, we demonstrate that a minimum of approximately 15 EMAs are needed to have a stable estimate of one's mean affect level. In contrast, stable estimates of affective variability appear to require between 20 and 25 EMAs. These minimums could be applied to existing data sets as ways of removing participants. That said, depending on the number of assessments being sent, the minimal number of EMAs necessary for a stable estimate is not necessarily identical to the percent of EMAs completed and both may be used. A final option for dealing with careless responders is to not explicitly exclude any participant but to remove the responses flagged as careless and use multilevel modeling methods. By using multilevel models, participants with fewer responses (and lower precision in their person-level estimate) will receive less weight when examining group-level fixed effects.

It is important to note that our EMA cleaning recommendations are based on results from a specific dataset with a specific sampling protocol. EMA research projects vary in the number of items, the complexity of the items, and the response formats available to participants. Because the generalizability of these cleaning protocols to all EMA data remains to be determined, these recommendations may be specific to our sampling context. Our recommendations are thus suggestive and not prescriptive. Apart from validating these specific cutoffs in other data sets, we believe these methods, and the accompanying R-package, EMAeval (see below), can help identify careless EMA responses across EMA studies. To that end, we strongly encourage EMA researchers to

Figure 17

Breakdown of the 281 Participants Who Had at Least One Response That Met Criteria for a Within EMA SD ≤ 20 OR TPI ≤ 1 s OR Percent Items at Mode ≥ 60%



Note. TPI = time per item; EMA = ecological momentary assessment. Each vertical line represents one participant, with the percentage of EMA that met our criteria indicated in various colors. Red identifies assessments with TPI ≤ 1 s, blue identifies *SD* ≤ 20, green identifies % items at mode ≥ 60%, and pink identifies two out of the three criteria were met, orange identifies all three criteria were met, and purple identifies no criteria met. See the online article for the color version of this figure.

apply these methods to their own EMA data, but determine whether these methods and cutoffs (or other cutoffs) identify careless responses in their own data. As such, we recommend researchers apply the cutoffs identified in this article only after inspecting their data-specific outputs from EMAeval.

Limitations

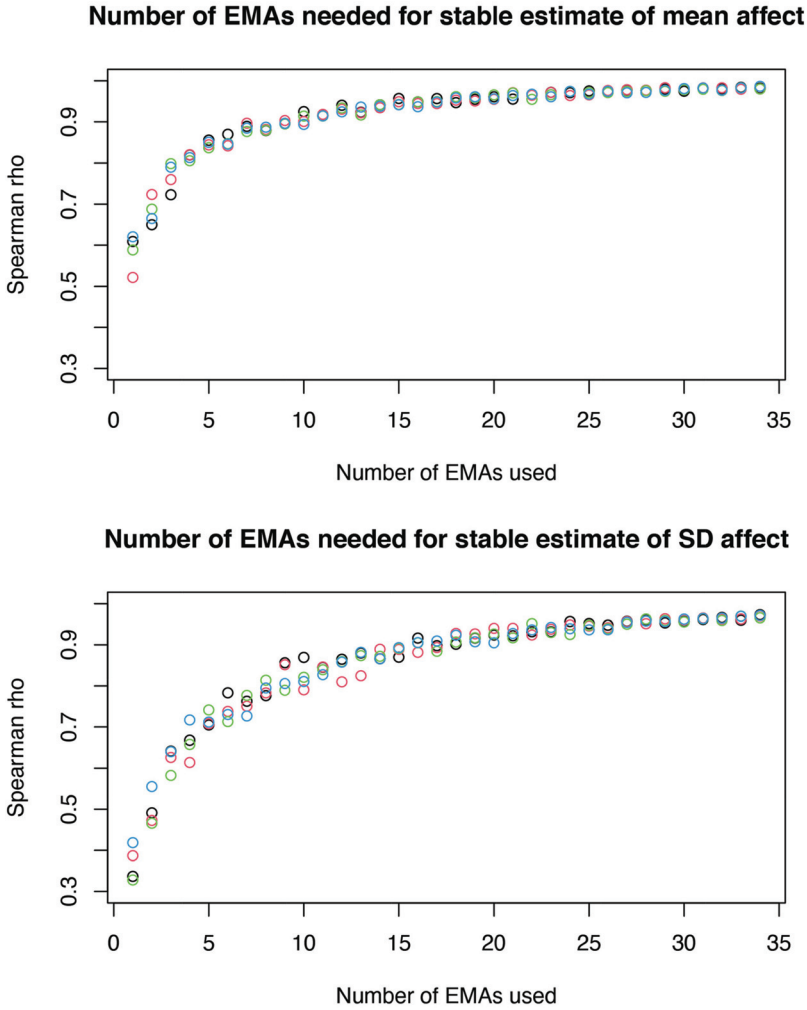
There are limitations to note. First, time per item was calculated based upon the time the assessment was started minus the time the assessment was completed, divided by the number of items. Thus, TPI reflects an average measure and is not specific to each item. Some web-based interfaces, such as Qualtrics, do not permit identifying the specific time per item unless each item is presented on a unique page. This approach is substantially more cumbersome for users. Yet, this more user-friendly approach limits the specificity with which we can infer individual TPI effects. Second, as we note above, we arrived at the *SD* ≤ 5 criterion not so much through a specific quantitative analysis, but via qualitative visualization. The distribution of the within EMA *SD* included two overlapping, but distinct distributions of within EMA *SD*s (see Figure 6). The presence of two such distributions suggests two distinct generative processes. Thus, while the *SD* criterion was effective at finding careless responses with very low *SD*, unlike the TPI criteria, it did not sample the entire response space evenly well. Third, our recommended cutoffs are broad recommendations based on an aggregated dataset. Clearly, some individuals may be able to complete valid EMAs quickly, while others may take longer to complete them and yet still provide careless responses. Without substantial data for that individual a priori, computing valid individualized metrics is impossible. Again, this limitation pointed us to recommending somewhat more lenient inclusion criteria. Lastly, and as noted above, only affect was assessed. Thus, the appropriate TPI may differ, depending on if cognitive, behavioral, personality, or other (e.g., somatic) constructs are being assessed.

R-Package to Identify Careless Responses and Responders

To accompany the analyses presented here and to promote data quality across EMA research, we created an R-package (called "EMAeval") so EMA researchers can apply the criteria outlined

Figure 18

Estimating the Number of EMAs Required for a Stable Estimate of Mean and Variability



Note. EMA = ecological momentary assessment. For each subject, we randomly sampled a subset of their EMA data (1, 2, . . . , 35 EMAs) and calculated the mean/*SD* of positive affect (PA) for that random sample. We also calculated the mean/*SD* of all EMAs for each subject as their "ground truth" estimate of mean PA and PA *SD*. We then calculated the across subject's rank-order correlation between the ground truth and the random sample. As can be seen in the figure, mean PA stabilizes around 15 EMAs, PA *SD* stabilizes around 20–25 EMAs. We performed this procedure four times and different colors represent each of the four iterations. Results are reliable across iterations. See the online article for the color version of this figure.

above for identifying careless responses and careless responders in their data. This package is available for researchers to download directly from our github page (https://github.com/manateelab/) with practice data or directly in R using the devtools::install_github ("manateelab/EMAeval-R-Package") function. There are several notable features to this package. First, the package includes a function that creates figures similar to those presented here. This is done so researchers can qualitatively determine their own cutoff values for TPI, within EMA item score *SD*, or proportion of items at the mode. Because different EMA studies will use a different number of items and those items may have different levels of complexity, the cutoff values described in our findings may need to be tweaked on a case-by-case basis, which is why visualization plays a vital role in data-cleaning. Second, users are provided with a reformatted data frame that organizes and summarizes the data used to develop plots and determine cutoffs. Third, once a researcher determines their own criteria for what classifies as a careless response, they will be able to run additional functions that flag responses or responders who have assessments that meet the defined cutoff criteria as well as any combination ("AND," "OR") of these criteria. This allows researchers the ability to customize their cutoff criteria based on their data type and relevant information that may inform the

desired use for these functions. This information will then be reformatted into an additional and separate data frame which will provide the participant identification number as well as the specific assessments that meet the researcher's criteria for a careless response. These functions will allow researchers to then perform their own independent analyses to confirm their desired cutoff criteria and notify problematic responders.

We anticipate this R-package can be used under two conditions. First, and most simply it can be used after data collection to clean data and remove careless responses and responders. Second, this package can be used while data is being collected to identify careless responders to limit poor quality data. Specifically, we envision creating a wrapper script to perform API calls to the survey interface being used (e.g., Qualtrics) to auto-download EMA data on a regular basis (e.g., weekly). These data would be auto-piped to the EMAeval package. The outputs from EMAeval will identify careless responses and importantly, subjects who may have begun to respond carelessly. Results can be sent to the experimenter to investigate further, or can initiate an automatically generated email to the participant, requesting that they consider their EMA responses more carefully. Given that data, is precious, expensive, and time-consuming to acquire, such an online system may improve study compliance and enhance data quality by bringing awareness to the research study. This may reduce careless responses, and yield more valid and reliable data. We are beginning to test such practices in our own laboratory.

Research using EMA has burgeoned in recent years. Here we demonstrate the importance of examining individual EMA responses for quality and we suggest three metrics: (1) how quickly an assessment has been completed, (2) whether the responses to an individual EMA have very low variance, and (3) if a high proportion of the responses to an individual EMA fall at the mode. We also demonstrate a method for determining if a research subject has sufficient number of responses or should be considered for removal. Lastly, we introduce an R-package for researchers to implement post hoc or online data evaluation and cleaning. We hope EMA researchers begin to employ these methods that we believe can improve the reliability and validity of EMA research going forward.

References

Anderson, M. (2019, June 13). *Mobile technology and home broadband*. Pew Research Center. https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019/

Beal, D. J., & Weiss, H. M. (2003). Methods of ecological momentary assessment in organizational research. *Organizational Research Methods*, *6*(4), 440–464. https://doi.org/10.1177/1094428103257361

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Bolger, N., & Zuckerman, A. (1995). A framework for studying personality in the stress process. *Journal of Personality and Social Psychology*, *69*(5), 890–902. https://doi.org/10.1037//0022-3514.69.5.890

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*(1), 579–616. https://doi.org/10.1146/annurev.psych.54.101601.145030

Cain, A. E., Depp, C. A., & Jeste, D. V. (2009). Ecological momentary assessment in aging research: A critical review. *Journal of Psychiatric Research*, *43*(11), 987–996. https://doi.org/10.1016/j.jpsychires.2009.01.014

Christensen, T. C., Barrett, L. F., Bliss-Moreau, E., Lebo, K., & Christensen, T. C. (2003). A practical guide to experience-sampling procedures. *Journal of Happiness Studies*, *4*(1), 53–78. https://doi.org/10.1023/A:1023609306024

Congdon, E., Mumford, J. A., Cohen, J. R., Galvan, A., Canli, T., & Poldrack, R. A. (2012). Measurement and reliability of response inhibition. *Frontiers in Psychology*, *3*, 37. https://doi.org/10.3389/fpsyg.2012.00037

Csikszentmihalyi, M. (2006). *The experience of psychopathology: Investigating mental disorders in their natural settings*. Cambridge University Press.

Csikszentmihalyi, M., & Larson, R. (1992). Validity and reliability of the experience sampling method. *The Journal of Nervous and Mental Disease, 175*(9), 526–536.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and reliability of the experience-sampling method. In M. Csikszentmihalyi (Ed.), *Flow and the foundations of positive psychology* (pp. 35–54). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9088-8_3

de Graaf, R., Bijl, R. V., Smit, F., Ravelli, A., & Vollebergh, W. A. M. (2000). Psychiatric and sociodemographic predictors of attrition in a longitudinal study: The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *American Journal of Epidemiology*, *152*(11), 1039–1047. https://doi.org/10.1093/aje/152.11.1039

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*(5), 478–491. https://doi.org/10.1038/s41562-019-0555-0

Ebner-Priemer, U. W., Kuo, J., Kleindienst, N., Welch, S. S., Reisch, T., Reinhard, I., Lieb, K., Linehan, M. M., & Bohus, M. (2007). State affective instability in borderline personality disorder assessed by ambulatory monitoring. *Psychological Medicine*, *37*(7), 961–970. https://doi.org/10.1017/S0033291706009706

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological Assessment*, *21*(4), 463–475. https://doi.org/10.1037/a0017075

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, *74*(4), 967–984. https://doi.org/10.1037/0022-3514.74.4.967

Goldberg, L. R. (2001). *The comparative validity of adult personality inventories: Applications of a consumer-testing framework*. Plenum Press.

Hufford, M. R. (2007). Special methodological challenges and opportunities in ecological momentary assessment. In A. A. Stone, S. Shiffman, A. A. Atienza, and L. Nebeling (Eds.), The science of real-time data capture: Self-reports in health research (pp. 54–75). Oxford University Press.

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

Jones, A., Remmerswaal, D., Verveer, I., Robinson, E., Franken, I. H. A., Wen, C. K. F., & Field, M. (2019). Compliance with ecological momentary assessment protocols in substance users: A meta-analysis. *Addiction*, *114*(4), 609–619. https://doi.org/10.1111/add.14503

Larsen, J. T., McGraw, A. P., & Cacioppo, J. T. (2001). Can people feel happy and sad at the same time? *Journal of Personality and Social Psychology*, *81*(4), 684–696.

McCabe, K. O., Mack, L., & Fleeson, W. (2012). A guide for data cleaning in experience sampling studies. In M. R. Mehl and T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 321–338). Guilford Press.

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

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of interindividual and intraindividual variability: A theoretical and empirical vindication of developmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development: Dialogues with lifespan psychology* (pp. 339–360). Springer. https://doi.org/10.1007/978-1-4615-0357-6_15

Russell, J. A. (1979). Affective space is bipolar. *Journal of Personality and Social Psychology*, *37*(3), 345–356. https://doi.org/10.1037/0022-3514.37.3.345

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*(3), 513–523. https://doi.org/10.1037/0022-3514.45.3.513

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., Iida, M., Gleason, M. E. J., Xu, J. H., & Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(1), E15–E23. https://doi.org/10.1073/pnas.1712277115

Stamatis, D. H. (2002). *Six sigma and beyond: Statistics and probability* (Vol. III). CRC Press.

Suls, J., Wan, C. K., & Blanchard, E. B. (1994). A multilevel data-analytic approach for evaluation of relationships between daily life stressors and symptomatology: Patients with irritable bowel syndrome. *Health Psychology*, *13*(2), 103–113. https://doi.org/10.1037/0278-6133.13.2.103

Van Berkel, N., Ferreira, D., & Kostakos, V. (2018). The experience sampling method on mobile devices. *ACM Computing Surveys*, *50*(6), 1–40. https://doi.org/10.1145/3123988

Vansimaeys, C., Zuber, M., Pitrat, B., Join-Lambert, C., Tamazyan, R., Farhat, W., & Bungener, C. (2017). Combining standard conventional measures and ecological momentary assessment of depression, anxiety and coping using smartphone application in minor stroke population: A longitudinal study protocol. *Frontiers in Psychology*, *8*, 1172. https://doi.org/10.3389/fpsyg.2017.01172

Villano, W. J., Otto, A. R., Ezie, C. E. C., Gillis, R., & Heller, A. S. (2020). Temporal dynamics of real-world emotion are more strongly linked to prediction error than outcome. *Journal of Experimental Psychology: General*, *149*(9), 1755–1766. https://doi.org/10.1037/xge0000740

Wang, L. P., Hamaker, E., & Bergeman, C. S. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*(4), 567–581. https://doi.org/10.1037/a0029317

Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form*. https://doi.org/10.17077/48vt-m4t2

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis. *Journal of Medical Internet Research*, *19*(4), e132. https://doi.org/10.2196/jmir.6641

Wheeler, L., & Reis, H. T. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality*, *59*(3), 339–354. https://doi.org/10.1111/j.1467-6494.1991.tb00252.x

Young, A. F., Powers, J. R., & Bell, S. L. (2006). Attrition in longitudinal studies: Who do you lose? *Australian and New Zealand Journal of Public Health*, *30*(4), 353–361. https://doi.org/10.1111/j.1467-842X.2006.tb00849.x

Received May 5, 2020
Revision received January 10, 2021
Accepted February 5, 2021 ■