Graduate Student Journal of Psychology
Fall 2025 - Vol. 25

Response Time to Detect Careless Responding and Its Relationship with and Prediction of Emotional Distress

Kristen E. Zentner & Seyma N. Yildirim-Erbasli
Department of Psychology, Concordia University of Edmonton

People experiencing emotional distress struggle with cognitive and motivational decline, which has been correlated with patterns of careless responding. Although several methods have been used to detect careless responses in emotionally distressed respondents, the response time has not been widely explored. The current study conducted secondary data analyses on a sample (N = 37,819) who completed the Depression Anxiety Stress Scale (DASS-42) in an online survey between 2017 and 2019. First, a response-time-based approach––a normative threshold method––was used to identify careless responding and examine its association with emotional distress using the DASS-42. Second, four machine learning models––decision tree (DT), random forest (RF), support vector machine (SVM), and naive Bayes (NB)––were trained on DASS-42 item responses and response times to predict emotional distress severity level. A significant correlation was found between the number of careless responses and subscale scores of anxiety and stress. In addition, Mann-Whitney U tests showed statistically significant differences between careless and careful responders in depression, anxiety, and stress. Regarding the machine learning models, SVM was found to be the best predictive model for classifying distressed people with an accuracy, sensitivity, and specificity exceeding 90%. Our results suggest that, in addition to survey responses, response time can identify careless responders and predict distressed responders.

Keywords: Response time, machine learning, psychological distress, careless responding

Computerized self-report measures have revolutionized the administration of psychological measures by offering researchers and clinicians a more efficient and convenient means of collecting data. However, the potential of computerized testing has yet to be fully realized. Additional data available in computerized testing (e.g., response time), which are not available through traditional pen-and-paper administration, offer significant contributions to the psychometric utility of these tests. Response time is recognized for its ability to provide a broader representation of responses, going beyond the responses themselves (van der Linden et al., 2010). It has considerable potential to reveal psychometrically relevant information, assess profile validity (i.e., the extent to which an individual's test score represents the true level of the trait or ability being measured), and develop increasingly precise and accurate adaptive testing methods.

Emotional Distress and Response Behavior

There is strong evidence to suggest the impact of emotion on performance on cognitive tasks in the cognitive psychology literature (e.g., Castaneda et al., 2008, 2011; Eysenck et al., 2007; Gross, 2015; Hubbard et al., 2016). Given that emotion and cognition interact to impact behavior, it is necessary for experimental researchers to consider how this interaction may impact the quality of self-report data in various contexts. For example, research indicates that individuals with greater levels of emotional distress have biases in their self-report data linked to this level of emotionality (Ashley & Shaughnessy, 2021; Conijn et al., 2020). Furthermore, researchers are often interested in measuring emotion itself. The Profile of Mood States (POMS; McNair et al., 1971), the Multiple Affect Adjective Checklist (MAACL; Zuckerman & Lubin, 1985), the Positive and Negative Affect Scale (PANAS-X; Watson & Clark, 1994), and the Depression Anxiety and Stress Scale (DASS-42; Antony et al., 1998) are frequently used self-reports that measure various aspects of emotional states. When a participant completes such self-reports, they are engaging in a cognitive task (i.e., completing the survey itself) that involves emotional content (i.e., the item content) which may interact with or influence the emotional or cognitive state of the participant, such as instigating a heightened emotional response or engagement in emotional regulation strategies (Castaneda et al., 2008; Gotlib & Joormann, 2010; Gross, 2015; Hubbard et al., 2016; Sun & Alkon, 2014). Despite recent research, gaps exist in our understanding of how emotion and cognition interact to impact response behavior on surveys.

Cognitive symptoms of emotional distress can impact data quality on self-report measures. Self-report surveys require a degree of effort to complete, and factors that compromise sustained effort may impact survey response styles. Several theoretical models have been applied to understand the relationship between emotion and cognition in response behavior. For example, cognitive bias literature suggests that mood-congruent information increases working memory in depressed individuals; that is, they have a bias to pay attention to information that reaffirms their depressive cognitions (Hubbard et al., 2016). However, drawing on cognitive behavioral theory, Ashley and Shaughnessy (2021) proposed that individuals with depression engage in avoidance behaviors when completing surveys, adopting strategies that

reduce the impact of distressing survey content by minimizing effort, attention, and time to completion. This calls into question whether self-reports of emotional distress are responded to with heightened attention due to the mood-congruent nature of the items, or with inattention (i.e., careless responding) due to concentration difficulties and avoidance behaviors characteristic of depression, anxiety, and stress symptoms. It is recognized that the cognitive impairments associated with depression, anxiety, and stress impact how these individuals respond to self-report measures, potentially threatening their profile validity and rendering scores inaccurate (Ashley & Shaughnessy, 2021; Conijn et al., 2020). Studying the patterns of response behavior on self-report measures in emotionally distressed respondents may clarify how their cognitive symptoms are impacting their survey responses and provide an effective way to assess profile validity.

Careless responding tends to be higher in populations with mental health concerns, with rates ranging from 6.0% (LePagea et al., 2001) to 12.6% (Conijn et al., 2015). Moreover, findings indicate that those with more severe psychopathology are more likely to show aberrant response styles (Conijn et al., 2015, 2018; Keeley et al., 2016; Wardenaar et al., 2015). Comorbid anxiety and depression appear to be associated with even greater aberrant responses because of the interacting effects of the two forms of psychological distress on working memory capacity (e.g., Beaudreau & O'Hara, 2009). The relationship between anxiety alone and response bias is less clear (Ferreri et al, 2011; Salthouse, 2012), possibly because those with anxiety do not consistently exhibit cognitive symptoms (Castaneda et al., 2008). Conijn et al. (2020) proposed and tested a theoretical explanation for careless responding in clinically depressed and anxious populations. They argued that cognitive symptoms of depression, specifically concentration, comprehension, and memory, limit cognitive abilities and make aberrant responses more likely (Hubbard et al., 2016). In testing their model, they found that cognitive symptoms mediate the relationship between depression or anxiety and response biases. Another study indicated that higher levels of anxiety, distress, and sadness were associated with inattention on surveys (Ashley & Shaughnessy, 2021).

When researchers fail to detect and report instances of careless responding, it impacts findings, jeopardizing the overall quality of knowledge production in the field. This threat to the psychometric properties of self-report measures has been reported and studied by many researchers, and its relevance to clinical settings has been explored (e.g., Cuijpers et al., 2010; Keeley et al., 2016; Tada et al., 2014). When self-report questionnaires are used diagnostically, clinicians base their clinical decision-making and diagnosis on information that may overestimate or underestimate symptom severity (Keeley et al., 2016). Given the findings from previous research and the psychometric utility of attending to careless responding, a clear understanding of careless response detection is needed.

Approaches to Understanding Response Behavior

There are a variety of approaches to detect careless responding (see Ward & Meade, 2023, for review). Proactive indices are those that place items within the survey itself to assess inattention, such as "Because I am paying attention, I will answer this question with '*Very little*'" (Ashley & Shaughnessy, 2021, p. 4). However, these single-item proactive indices provide little contextual information about the pattern of careless responding throughout an entire survey. Reactive indices are those that flag inattentive responders through detection of careless response styles during data cleaning and analysis phases, such as longstring detection (i.e., the selection of the same response option for several consecutive responses), and participant-specific reliability (i.e., the consistency of a participant's responses on items measuring the same trait; Ashley & Shaughnessy, 2021). Curran (2016) noted that the use of a single response style approach is insufficient to detect careless responders because they may have response style patterns that are detectable with some approaches but not others. For example, participants who use a longstring response style would have high participant-specific reliability and would not be identified by detection approaches designed to detect even-odd response styles where participants select extreme ends of a scale (Meade & Craig, 2012). To account for this diversity in response styles, researchers have suggested that the use of multiple detection approaches is necessary to identify careless responders (Ashley & Shaughnessy, 2021; Curran, 2016).

In their review of careless responding, Ward and Meade (2023) suggest that extensive screening methods may be necessary when analyzing large datasets or with populations that are more likely to engage in careless responding, such as emotionally distressed individuals. For example, Ashley and Shaughnessy (2021) found that proactive items and short survey response time were associated with negative emotional states (i.e., sadness, anxiety, distress) while other detection methods (e.g., longstring, participant-specific reliability) were not.

Response Time Approaches

Response time approaches, which operate on the assumption that careless responders will have unreasonably

rapid survey response times consistent with the motivation to finish the survey quickly, have received considerable review within the literature (Ashley & Shaughnessy, 2021; Curran, 2016; Jones et al., 2022; Ward & Meade, 2022). Researchers have identified that response time has the potential as a detection method on computerized tests because it is more difficult to manipulate than other methods (e.g., longstring, participant-specific reliability; Curran, 2016). For example, careless responders who wish to appear as careful may be motivated to provide response patterns that mimic normal responses (e.g., selecting responses consistently at one end of the scale with enough variability to avoid pattern detection of longstring and even-odd styles). Due to their desire to finish the survey quickly, however, they will likely still have shorter response times than careful responders, as demonstrated by Schnipke and Scrams (1997).

Careless and careful responders exhibit distinct distributions in response time, with an initial spike in response time distribution attributed to the former. Curran (2016) used simulated data to explore distributions of response time in careless and careful responders and highlighted consideration of Type I (i.e., falsely identifying a careful responder as careless) and Type II errors (i.e., falsely identifying a careless responder as careful) in establishing cut-off scores due to the significant overlap between distributions. Researchers have used various calculations to determine cut-off scores, including 1.5 quartiles above or below the median (Funke, 2016), two standard deviations above or below the mean (Heerwegh, 2003), one standard deviation above or below the mean (Ashley & Shaughnessy, 2021), and various percentiles (e.g., first percentile, fifth percentile; Gummer & Roßmann, 2015; Harms et al., 2017). Some researchers have explored cut-off scores on an individual item level. For example, a 2-second-per-item cut-off score is considered a conservative approach, limiting Type II errors at the cost of missing some careless responders (Bowling et al., 2016; Huang et al., 2012).

Several methods have been used to determine response time thresholds for individual items, including a two-state mixture model (Schnipke & Scrams, 1997), surface-level characteristics of items (i.e., character count; Wise & Kong, 2005), and visual inspection of the response time-frequency distribution (Wise, 2006). These three approaches tend to identify similar item response time thresholds (Kong et al., 2007). Considering that items vary in the amount of text or how mentally taxing they are, what may be classified as rapid responding varies by the item. This implies that greater accuracy of careless response detection may be gained by identifying a normative threshold specific to each item, rather than using a general response time threshold applied to all questions as has been done previously (Huang et al., 2012; Wise & Ma, 2012). The normative threshold approach involves the use of response time cut-off scores to identify individual items that are responded to carelessly (Wise & Ma, 2012).

Response time approaches have been of limited utility in psychological research compared to other careless responding detection methods because this metric is typically available at a page or survey level (Ashley & Shaughnessy, 2021; Ward & Meade, 2023). However, with item response time available at the item level, researchers and clinicians can gain precision in identifying response time patterns of careless responders in emotional distress. The greatest contribution of item response time to the careless responding literature might be its potential to detect careless responders regardless of the responders' specific response style (e.g., long string), which gives researchers detailed information on survey response patterns while blocking respondents' attempts to mask their response style (Curran, 2016).

Computerized surveys that provide access to response time at the item level also create opportunities for more sophisticated analyses. For example, computerized adaptive testing uses an algorithm to select items based on previous responses to gain efficiency and precision in the measurement of the ability or trait with the administration of fewer items (Wise, 2020). Companies that produce widely used psychological measures are increasingly moving to computerized adaptive formats for reduced testing time and ease of administration and scoring (Forbey et al., 2012). Wise (2020) notes that traditional approaches to computerized adaptive testing in education, which use only item difficulty in adaptation algorithms, could be expanded to use behavioral measures such as response time to improve the precision and accuracy of measurement. Computerized adaptive testing relies on advanced modeling, such as machine learning approaches, which can be used to examine response and response time patterns to predict emotional distress.

Machine Learning Approaches to Predict Emotional Distress

Several researchers have used machine learning methods with item response data to explore patterns that inform psychologists about how emotionally distressed individuals respond to self-report surveys. This line of research stems from machine learning's ability to capture subtle patterns not evident through traditional approaches and its effectiveness in handling the complex interactions and dependencies among variables, which are likely to be common in psychological assessment data collected by surveys such as the DASS. For

example, Budiyanto et al. (2019) used text mining from social media posts to measure depression and anxiety using a closed-loop machine learning approach, with an NB algorithm as a training process and the DASS-21 parameters as a learning process. From this approach, Budiyanto et al. (2019) demonstrated the usefulness of machine learning methods to collect novel information about emotional distress. Other studies have found that various machine learning models are useful in predicting depression, anxiety, and stress from the DASS-42 and DASS-21, with some machine learning models showing greater accuracy and efficiency than others (Kumar et al., 2020; Priya et al., 2020; Srinath et al., 2022).

Kumar et al. (2020) predicted five severity levels of emotional distress by investigating eight machine learning algorithms trained on DASS-42 item responses, and then the same methods were applied to a second DASS-21 dataset. The results showed that these models could be used to predict emotional distress, with accuracy rates between 96.02% and 97.48% for the subscales (Kumar et al., 2020). Priya et al. (2020) applied five machine learning models trained on item response data, including DT, RF, NB, SVM, and *K*-nearest neighbor, to predict depression, anxiety, and stress levels from a sample of DASS-21 data. They found that the RF classifier demonstrated the best performance, with accuracy rates between 71.4% and 79.8% for DASS subscales. Srinath et al. (2022) compared SVM and logistic regression using parameter tuning to predict depression, anxiety, and stress from DASS-42 item response data. They found that logistic regression had the highest performance, with an accuracy of 98.15% for depression, 98.05% for anxiety, and 98.45% for stress. In recent years, researchers have begun to explore the utility of machine learning models using response time on behavioral tasks (i.e., perceptual matching task) and using neuroimaging and physiological data (i.e., Magnetic Resonance Imaging; Liu et al., 2022). These studies suggest that machine learning has predictive potential within clinical and counseling psychology. Moreover, Priya et al. (2020) noted that the sensitivity and specificity afforded by machine learning models make these approaches particularly helpful within healthcare contexts.

Current Research

Considering how often self-report is used to measure, research, and reduce impairment from negative emotional states such as depression, anxiety, and stress, further exploration of careless response identification is needed to enhance data quality and continue to elucidate the impact of emotion on cognitive tasks. While response time has been used as a measure of response behavior (e.g., Kong et al., 2007), more study of careless response detection within emotionally distressed populations is needed to determine effective ways to identify and deal with potentially invalid data. In addition, the use of machine learning approaches can facilitate the identification of complex patterns and relationships that may not be apparent through traditional statistical methods. There is a recent body of literature on the use of machine learning approaches to predict emotional distress using item response data (e.g., Srinath et al., 2022). However, this study proposes incorporating response time in addition to item responses, as response time can provide insights beyond the responses themselves (van der Linden et al., 2010). The present exploratory study aims to address these gaps in the literature by examining the relationship between careless responding and emotional distress, as well as exploring the potential utility of incorporating response time in machine learning models for predicting emotional distress. In this paper, the following research questions were investigated:

1. Is there an association between careless responding and emotional distress (i.e., depression, anxiety, and stress)?
2. Can machine learning models identify emotionally distressed people using item responses and response time?

Methods

The DASS-42 is a well-established measure of emotional distress with 42 items such as, "I felt that life was meaningless" in the depression subscale, "I was aware of dryness in my mouth" in the anxiety subscale, and "I found that I was very irritable" in the stress subscale on a scale of 0 (Did not apply to me at all) to 3 (Applied to me very much or most of the time; Lovibond & Lovibond, 1995). The subscales assess depression ($\alpha = .97$), anxiety ($\alpha = .92$), and stress ($\alpha = .95$) as separate constructs with 14 items each, and each demonstrates high internal consistency (Antony et al., 1998).

To address the research questions in this study, the dataset was pulled from the Open Source Psychometrics Project (2019), which offers public datasets. The survey was open for anyone to complete, meaning the sample may include a mixture of clinical and non-clinical populations. Participants received only their personalized results in return for their participation. The dataset was pre-cleaned upon download–– negative response times were recoded to missing values, and milliseconds were transformed into seconds. Data analysis was conducted using jamovi and R programming languages. The sample consisted of 37,819 participants who completed the online survey on a scale of 1 to 4 between 2017 and 2019,

aged 13 to 79 years (M = 23.39, SD = 8.57). Sociodemographic characteristics of the sample are displayed in Table 1.

Question 1: Association between Careless Responding and Emotional Distress

Figure 1 encapsulates the methodology employed to address our primary research question. To answer our first research question, we identified careless responses by calculating a normative threshold based on Wise and Ma's (2012) response time approach. Second, we classified participants as careless or careful upon considering the total number of careless responses they exhibited. Third, we examined the association between the number of careless responses and emotional distress scores. Finally, we conducted Mann-Whitney U tests to compare emotional distress scores between careless and careful responders.

The normative threshold for an item is calculated as "a percentage of the elapsed time between when the item is displayed and the mean of the response time distribution for the item, up to a maximum threshold value of ten seconds" (Wise & Ma, 2012, p. 9). For example, if an item takes a mean of 60 seconds for participants to complete, the 10% normative threshold (i.e., NT10) would be six seconds. Various normative thresholds can be compared to determine the cut-off that yields the greatest accuracy for identifying careless responders (Wise & Ma, 2012). We chose to use the 20% normative threshold (NT20) based on literature showing that this cut-off is appropriate for low-stakes environments (Rios & Soland, 2021) and due to the positively skewed distribution of DASS-42 scores in the dataset. To calculate the NT20, the mean response time was calculated for each item, and 20% of the average response time served as the NT20 cut-off score. We recoded careless responses (i.e., responses below the NT20 cut-off) as missing values. The reason behind this is that previous studies have pointed out how the presence of careless responses in the dataset can introduce bias into estimates of item and person parameters (e.g., Guo et al., 2016).

Various cut scores have been used to classify careful and careless responders (i.e., demonstrating a substantial number of careless responses throughout the survey; e.g., Wise & Kong, 2005). The purpose of the tool and the sample under study are crucial factors in determining appropriate cut scores. For example, a cut score of 20% was used in a low-stakes assessment (e.g., Wise & Kong, 2005). When a survey is used diagnostically for clinical purposes, prudent clinicians must be confident that the data are not impacted by careless responding, while putting more weight on other data sources (i.e., interviews) if the survey data have questionable validity (American Psychological Association, 2020). Stated another way, increasing false positives (classifying careful responders as careless) may be necessary for evaluating the validity of clinically relevant data. Given that the survey in the current study measures clinically relevant variables, a cut score of 10% was chosen. By setting the threshold at this level, we aimed to be inclusive enough to detect individuals who may exhibit a notable pattern of inattentive responses across the survey items. At the same time, the 10% threshold is chosen to avoid categorizing individuals as careless responders when they may, in fact, be providing thoughtful and considered answers to the survey questions. Participants who showed between 0 and 4 careless responses within the DASS-42 items were classified as careful responders (n = 37,025), and participants who showed between 5 and 20 careless responses were classified as careless responders (n = 551). Participants with 20 or more careless responses were considered extremely careless responders (n = 243) and were excluded from analyses because their survey scores would have been significantly biased by the severity of their careless responding. This is consistent with previous literature that uses a 50% careless response rate as a cut-off for removal from the dataset (Arias et al., 2020; Curran, 2016).

Given that data from carelessly responded items is invalid and introduces bias (e.g., Guo et al., 2016), it is necessary to recode these responses as missing values and compute scores accordingly. Therefore, after identifying careless responses, DASS-42 subscale scores were adjusted to represent only carefully responded items by recoding these responses as missing values and calculating the adjusted total scores. The adjusted subscale score is the sum of the scores for carefully responded items divided by the maximum possible total score for those items. We presented the adjusted scores as percentages to facilitate the interpretation of the results. For example, if a participant had careless responses on three items on the depression subscale (14 items), their depression subscale score would be the sum of their scores on the remaining carefully responded 11 items, with a score range of 11 (11 × 1 point) to 44 (11 × 4 points). If the participant scored 39 on these 11 carefully responded depression items, their adjusted subscale score in percentage would be calculated as follows: $39/44 = 0.87 \times 100 = 87$.

A Spearman correlation analysis was used to determine whether an association exists between the number of careless responses and emotional distress scores. Next, Mann-Whitney U tests were used to compare DASS-42 subscale scores between participants classified as careless and careful.

Question 2: Predicting Emotional Distress with Ma-

chine Learning

Figure 2 encapsulates the methodology employed to address our second research question. To answer our second research question, we used DASS-42 data consisting of item responses and response times of 37,819 respondents. Responders were classified into five severity levels on an ordinal scale for depression, anxiety, and stress based on their scores using the guide for severity levels of emotional distress in DASS-42 (see Table 2). Total scores were calculated on a 0 to 3 scale by subtracting one from each response, as the initial dataset included responses ranging from 1 to 4. For each emotional distress, we used 80% of the dataset for training and 20% for testing. Validation was conducted within the training process through 10-fold cross-validation (i.e., the number of groups that the dataset is randomly split into) and a random search for hyperparameter optimization.

We trained four machine learning algorithms–DT, RF, SVM, and NB (James et al., 2017)–to predict the severity levels of responders for each emotional distress. DTs split data into progressively smaller subsets based on selected features to form a simple, interpretable tree structure. RFs enhance this approach by combining multiple DTs built from random subsets of data and features to improve accuracy and reduce overfitting. NB applies Bayes' theorem under the assumption that all features are independent to generate probability-based classifications. SVMs identify the optimal hyperplane with the maximum margin to separate data points for predictive performance. After the training, we evaluated the performance of the trained models using the test sets and reported standard classification metrics (i.e., sensitivity, specificity, and accuracy).

Results

Careless Responding and Emotional Distress

There were statistically significant correlations between the number of careless responses and subscales of anxiety ($r_s(37{,}574) = .03$, $p < .001$) and stress ($r_s(37{,}574) = .02$, $p = .001$), but not the subscale of depression ($r_s(37{,}574) = -.001$, $p = .836$).

In terms of the subscale of depression, careless responders ($Mdn = 67.3$) had higher scores than careful responders ($Mdn = 62.5$), and the difference between careless and careful responders was statistically significant ($U = 9.58e+6$, $p = .01$, $r = .06$; see Figure 3). Similarly, a statistically significant difference was found between groups on the anxiety subscale ($U = 9.00e+6$, $p < .001$, $r = 0.12$) with careless responders ($Mdn = 57.1$) scoring higher on anxiety than careful responders ($Mdn = 51.8$). Finally, higher stress subscale scores were found among careless responders ($Mdn = 65.9$) than careful responders ($Mdn = 62.5$) with a statistically significant difference ($U = 9.62e+6$, $p = .02$, $r = .06$). In summary, careless responders had higher scores of depression, anxiety, and stress than careful responders.

Predicting Emotional Distress

Table 3 shows single classification metrics of the four machine learning models for five different severity levels of each emotional distress. Even though machine learning models showed roughly similar performance, DT, RF, and NB yielded inconsistent sensitivity values for mild, moderate, and severe levels. In terms of the depression subscale, classification metrics for almost every level exceeded the acceptable threshold of 70% or the optimal threshold of 80% with several exceptions of sensitivity values being less than the acceptable threshold. Overall, the SVM outperformed the other models with classification metrics exceeding the optimal threshold for each level. Regarding the anxiety subscale (see Table 3), with the exception of SVM, the other models showed mixed and extremely low sensitivity values for the levels of mild, moderate, and severe, while others exceeded the optimal threshold of 80%. Similar to depression, SVM surpassed the other models with classification metrics above the optimal threshold for each level. For the stress subscale (see Table 3), DT, RF, and NB showed mixed and low sensitivity values for the levels of mild, moderate, and severe. Other measures went over the optimal threshold of 80% and, in particular, SVMs dominated the other models in terms of sensitivity and specificity, which were either 100% or very close.

We also macro-averaged the single metrics by calculating the averages of sensitivity and specificity values over severity levels (see Figure 4). Based on the macro-averaging, four machine learning models for three types of emotional distress showed similar results in terms of specificity values being larger than 90%, exceeding the optimal threshold, whereas they showed mixed results for sensitivity. Only SVM exceeded the optimal threshold, with sensitivity values being larger than 90% across all three types of emotional distress. These results reveal that true negatives can be predicted with optimal sensitivity and specificity by all four classification models. However, both true positives and true negatives can be predicted with optimal sensitivity and specificity by only SVM. In addition, among all four machine learning methods, SVM had the highest accuracy of classification compared to the other methods across all three emotional distress, followed by RF, DT, and NB (see Figure 5).

Discussion

The current exploratory study contributes to the literature by investigating the relationship between careless responding and self-reported depression, anxiety, and stress using response behavior information (i.e., response time). Additionally, the inclusion of item response time in machine learning models was explored for predicting depression, anxiety, and stress severity levels. These findings are germane to researchers who use computerized self-report measures, as response time can aid in the identification of careless responders who bias datasets and invalidate individual testing profiles. Furthermore, machine learning models can efficiently predict the severity level of emotional distress by taking not only item responses but also response time patterns into account.

Response time measures at the page and survey levels have been shown to have limited utility compared to other measures of careless responding (Ashley & Shaughnessy, 2021; Ward & Meade, 2022); however, in line with previous research, the current study illustrated the usefulness of behavioral response data at the individual item level. Since careless responders are assumed to have rapid response times, which is consistent with their motivations to finish the survey quickly, the item response time is expected to catch careless responders regardless of their response style (Curran, 2016). This careless response detection method has potential utility for any self-report dataset with item-level response time.

In addition to the detection of careless responders for data quality purposes, the findings of the current study indicate that item response behavior provides clinically relevant information: emotional distress is correlated with a behavioral measure at an item-specific level. Researchers have drawn theoretical links between negative emotional states (i.e., depression, anxiety, and stress) and careless responding behavior (Ashley & Shaughnessy, 2021; Conjin et al., 2020). The cognitive and emotional characteristics of emotional distress are theorized as the mechanism explaining high rates of careless responding in emotionally distressed individuals (Ashley & Shaughnessy, 2021; Conjin et al., 2020). However, additional research is needed to clarify the links between careless responding and depression, anxiety, and stress. For example, some theories suggest that depressed individuals have heightened attention towards mood-congruent stimuli (e.g., survey items), while others suggest that avoidance of mood-congruent stimuli is typical in depressed individuals (Hubbard et al., 2016). The association between item response time and emotional distress found in the current study suggests that item response time can be a novel and precise approach to testing theories of emotional distress by unpacking the patterns of careless responding associated with specific emotional states (Castaneda et al., 2008, 2011; Snyder et al., 2015a, 2015b).

Finally, the current study found that machine learning trained on DASS-42 item response and response time can be another approach to predicting the severity of emotional distress, which corroborates previous research (e.g., Kumar et al., 2020; Priya et al., 2020; Srinath et al., 2022). While achieving 100% accuracy may not be feasible or necessary, the goal of using machine learning in this paper was not necessarily to outperform simple arithmetic summation, but rather to show the potential of a data-driven approach to analyze and predict emotional distress by considering both item response and response time. However, using machine learning based on item responses and response times can offer several advantages over the simple arithmetic summation of response data. First, it can allow for capturing subtle patterns that may not be evident through simple arithmetic summation. Second, machine learning techniques can handle complex interactions and dependencies among variables, which are very likely to be present in psychological assessment data such as DASS. Furthermore, these models have the potential to generalize to new datasets and populations, provided that they are trained on diverse and representative samples. This could enhance the applicability of the predictive models across different settings and populations, ultimately improving their utility in research contexts or other applied settings (e.g., healthcare).

Implications and Recommendations

Psychological measurement is in a period of advancement, with computerized testing affording researchers new ways to collect and interpret data. Item response time can be used to identify careless responding, and it has the potential to untangle the psychological mechanisms behind carelessness, particularly in those experiencing emotional distress. The relationship between careless responding and emotional distress, as well as the prediction of emotional states considering both item response and response time, can have important implications for cognitive psychology researchers. Thus, the current study has several research, psychometric, and applied implications and future directions for consideration.

First, researchers who use online self-report surveys can use the normative threshold method to flag careless responders, which allows the researchers to identify if these responders and their responses are influencing the data and obscuring important findings. This may be of particular

importance among research pools in which the participants are receiving incentives to complete a survey. In such cases, participants may be motivated to finish the survey as quickly as possible to earn the incentive, leading to careless responses. This can result in low-quality data and invalid conclusions drawn from the survey results. If researchers are specifically using measures of emotional distress, such as the DASS-42, response time may help to identify those who have higher scores due to careless responding versus those who are genuinely emotionally distressed.

Second, the use of response time methods has clinical utility in assessing the profile validity of individuals who may be carelessly responding due to their emotional distress. Many commonly used psychological tests, including the DASS-42, do not include profile validity measures. One reason for this is that additional profile validity scales, such as positive impression management or defensiveness, add items to measures that are otherwise constructed to be as short and efficient as possible. Using response time to assess profile validity adds no additional items to these measures. Similar to profile validity measures, response time may also provide clinically relevant information to improve our understanding of the cognitive impairments that accompany emotional states. For example, item response time may help ascertain whether an emotionally distressed respondent tends to respond carelessly to avoid mood-congruent information and when they are biased to attend more carefully to mood-congruent information (Ashley & Shaughnessy, 2021; Hubbard et al., 2016). This is valuable information for treatment planning because useful interventions may vary based on whether a client over-attends to negative information (i.e., ruminates) or employs avoidance behavior. Thus, in terms of the first research question addressed in this study, understanding the relationship between careless responding and emotional distress can provide insight into the cognitive processes underlying emotional states.

Regarding the second research question, machine learning models can be utilized to analyze hidden patterns in both item response and response time for predicting self-reported measures of emotional states. Considering the current research that has demonstrated an association between emotional distress and response time, it is essential to incorporate response time into the assessment of emotional distress severity rather than relying solely on the arithmetic summation of responses. By doing so, we can achieve a better understanding of emotional distress.

Machine learning models that incorporate behavioral data, such as item response time, also have practical applications for adaptive testing. Wise (2020) suggested that the inclusion of item response data in adaptive testing allows test developers to provide a measure of attention to be considered in profile validity. With real-time monitoring of attention using item response data, developers can intervene to re-engage a respondent who is exhibiting careless responding. For example, if a respondent has several consecutive careless responses, a prompt may appear to remind them to carefully attend to each item. In educational contexts, response time has been included in adaptive testing models, but behavioral measures have not been widely used in computerized adaptive testing of personality and psychopathology.

Limitations and Future Research

There are several limitations in the current study. First, the use of large datasets is at higher risk of finding spurious correlations between variables. The current findings are situated within the theoretical and research literature supporting the assertion that cognitive symptoms of emotional distress impact response behavior, and thus provide greater confidence in the validity of the findings. Second, due to the nature of survey data, the direction of causation between careless responding and emotional distress cannot be confirmed. Other factors that were not studied in the current research (e.g., education, age, and formal diagnoses) may be confounding the relationship between emotional distress and careless responding. Similarly, the normative response method used to identify careless responses may be impacted by confounding variables, such as technical issues experienced by respondents, item wording, and item valence. Additional research is needed to understand how patterns of slow response time may be related to emotional distress due to low processing speed and poor concentration, and to differentiate these responses from slow, careful responders. The literature would benefit from a direct comparison of the normative threshold method with other detection methods (e.g., longstring).

There are also several limitations related to the second research question. First, there was a positive skew in our dataset with an overrepresentation of emotionally distressed responders. Kumar et al. (2020) noted the problem of determining the best predictive model when data is imbalanced between classification categories. Given the potential increase in computation time and considering the primary focus of our paper, we opted not to perform class balancing techniques. Future research can study different sampling methods, such as undersampling, oversampling, and ROSE techniques, to address the class imbalance. Second, the algorithm used in the machine learning model and its parameters may

have impacted model performance. The current study used DT, RF, NB, and SVM algorithms, but future research can study other algorithms not included here. Third, machine learning models may make biased predictions for groups belonging to different demographic categories, such as race, gender, and age. Future research must consider these demographic variables to understand the generalizability of the model to different populations. Finally, to mitigate concerns of circularity, we split the data into separate training and validation sets, ensuring that predictive analyses were conducted on an independent dataset. Future research could further strengthen the validation of response-time-based indicators by examining their predictive utility using additional independent outcome measures.

The current study stimulates several future research directions. First, future research may specify which emotional states (e.g., depression, anxiety, excitement, and boredom) and demographics (e.g., age and gender) are associated with higher levels of carelessness for enhanced psychometric accuracy. This approach especially benefits individuals who score low on the DASS-42 due to poor insight into or masking of their emotional state but whose cognitive impairment is indicated by a behavioral measure, such as response time. Second, the identification of patterns in the data that are indicative of certain levels of emotional distress, which could help to improve the diagnostic accuracy of the DASS-42 scale, should be explored. For example, researchers can investigate whether a client's carelessness increases, ebbs and flows, or has a consistent rate throughout the survey. Thirdly, the normative response time approach may be applied within cognitive psychology and emotion regulation research to explore how different emotion regulation strategies impact cognitive processes and to compare response time cut scores amongst different populations. Finally, more research is needed to explore how machine learning models that include response time can be incorporated effectively into clinical and psychological assessments, such as adaptive testing and wellness-oriented smartphone applications.

Conclusion

Careless responding is a significant source of bias in online self-report surveys–a common data collection method in the field of psychology. The normative threshold method is an important approach for researchers to identify careless responders, and it circumvents the limitations of other approaches for detecting response bias. The normative threshold method also offers a novel behavioral measure for studying the impact of emotional distress on cognition. While all researchers using self-report measures hold responsibility and accountability for collecting valid and reliable data, those conducting research on emotional distress face additional validity threats because of the relationship between emotional distress and careless responding. The current study provides preliminary evidence that incorporating normative response thresholds into routine data cleaning practices and machine learning models may enhance the accuracy with which psychological researchers can describe and predict emotional distress.

References

American Psychological Association. (2020). *APA Guidelines for psychological assessment and evaluation.* APA Task Force on Psychological Assessment and Evaluation Guidelines. https://www.apa.org/about/policy/guidelines-psychological-assessment-evaluation.pdf

Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., & Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. *Psychological Assessment, 10*(2), 176–181. https://doi.org/10.1037/1040-3590.10.2.176

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods, 52*(6), 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

Ashley, M., & Shaughnessy, K. (2021). Predicting insufficient effort responding: The relation between negative thoughts, emotions, and online survey responses. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement, 55*(3), 198–209 https://doi.org/10.1037/cbs0000308

Beaudreau, S. A., & O'Hara, R. (2009). The association of anxiety and depressive symptoms with cognitive performance in community-dwelling older adults. *Psychology and Aging, 24*(2), 507–512. https://doi.org/10.1037/a0016035

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111*(2), 218–

229. https://doi.org/10.1037/pspp0000085

Budiyanto, S., Sihombing, H. C., & Rahayu, I. M. F. (2019). Depression and anxiety detection through the closed-loop method using DASS-21. *Telkomnika, 17*(4), 2087–2097. https://doi.org/10.12928/TELKOMNIKA.v17i4.12619

Castaneda, A. E., Tuulio-Henriksson, A., Marttunen, M., Suvisaari, J., & Lönnqvist, J. (2008). A review on cognitive impairments in depressive and anxiety disorders with a focus on young adults. *Journal of Affective Disorders, 106*(1-2), 1–27. https://doi.org/10.1016/j.jad.2007.06.006

Castaneda, A. E., Suvisaari, J., Marttunen, M., Perälä, J., Saarni, S. I., Aalto-Setälä, T., Lönnqvist, J., & Tuulio-Henriksson, A. (2011). Cognitive functioning in a population-based sample of young adults with anxiety disorders. *European Psychiatry, 26*(6), 346–353. https://doi.org/10.1016/j.eurpsy.2009.11.006

Conijn, J. M., Emons, W. H. M., De Jong, K., & Sijtsma, K. (2015). Detecting and explaining aberrant responding to the Outcome Questionnaire-45. *Assessment, 22*(4), 513–524. https://doi.org/10.1177/1073191114560882

Conijn, J. M., Emons, W. H. M., Page, B. F., Sijtsma, K., Van der Does, W., Carlier, I. V. E., & Giltay, E. J. (2018). Response inconsistency of patient-reported symptoms as a predictor of discrepancy between patient and clinician-reported depression severity. *Assessment, 25*(7), 917–928. https://doi.org/10.1177/1073191116666949

Conijn, J. M., van der Ark, L. A., & Spinhoven, P. (2020). Satisficing in mental health care patients: The effect of cognitive symptoms on self-report data quality. *Assessment, 27*(1), 178–193. https://doi.org/10.1177/1073191117714557

Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review, 30*(6), 768–778. https://doi.org/10.1016/j.cpr.2010.06.001

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

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion, 7*(2), 336–353. https://doi.org/10.1037/1528-3542.7.2.336

Ferreri, F., Lapp, L. K., & Peretti, C. S. (2011). Current research on cognitive aspects of anxiety disorders. *Current Opinion in Psychiatry, 24*(1), 49–54. https://doi.org/10.1097/YCO.0b013e32833f5585

Forbey, J. D., Ben-Porath, Y. S., & Arbisi, P. A. (2012). The MMPI–2 computerized adaptive version (MMPI–2–CA) in a Veterans Administration medical outpatient facility. *Psychological Assessment, 24*(3), 628–639. https://doi.org/10.1037/a0026509

Funke, F. (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review, 34*(2), 244–254. https://doi.org/10.1177/0894439315575477

Gotlib, I. H., & Joormann, J. (2010). Cognition and depression: Current status and future directions. *Annual Review of Clinical Psychology, 6*, 285–312. https://doi.org/10.1146/annurev.clinpsy.121208.131305

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry, 26*(1), 1–26. https://doi.org/10.1080/1047840X.2014.940781

Gummer, T., & Roßmann, J. (2015). Explaining interview duration in web surveys: A multilevel approach. *Social Science Computer Review, 33*(2), 217–234. https://doi.org/10.1177/0894439314533479

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173–183. https://doi.org/10.1080/08957347.2016.1171766

Harms, C., Jackel, L., & Montag, C. (2017). Reliability and completion speed in online questionnaires under consideration of personality. *Personality and Individual Differences, 111*, 281–290. https://doi.org/10.1016/j.paid.2017.02.015

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review, 21*(3), 360–373. https://doi.org/10.1177/0894439303253985

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*, 372–386. http://doi.org/10.1080/02796015.2005.12086292

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

Hubbard, N. A., Hutchison, J. L., Turner, M., Montroy, J., Bowles, R. P., & Rypma, B. (2016). Depressive thoughts limit working memory capacity in dysphoria. *Cognition and Emotion, 30*, 193–209. https://doi.org/10.1080/02699931.2014.991694

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. New York: Springer.

Jones, A., Earnest, J., Adam, M., Clarke, R., Yates, J., & Pennington, C. R. (2022). Careless responding in crowdsourced alcohol research: A systematic review and meta-analysis of practices and prevalence. *Experimental and Clinical Psychopharmacology, 30*(4), 381–399. https://doi.org/10.1037/pha0000546

Keeley, J. W., Webb, C., Peterson, D., Roussin, L., & Flanagan, E. H. (2016). Development of a response inconsistency scale for the personality inventory for DSM–5. *Journal of Personality Assessment, 98*(4), 351–359. https://doi.org/10.1080/00223891.2016.1158719

Kilgus, S. P., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the Social and Academic Behavior Risk Screener for elementary grades. *School Psychology Quarterly, 28*(3), 210–226. https://doi.org/10.1037/spq0000024

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Kumar, P., Garg, S., & Garg, A. (2020). Assessment of anxiety, depression and stress using machine learning models. *Procedia Computer Science, 171*, 1989–1998. https://doi.org/10.1016/j.procs.2020.04.213

LePagea, J. P., Mogge, N. L., & Sharpe, W. R. (2001). Validity rates of the MMPI-2 and PAI in a rural inpatient psychiatric facility. *Assessment, 8*(1), 67–74. https://doi.org/10.1177/107319110100800106

Liu, Y. S., Song, Y., Lee, N. A., Bennett, D. M., Button, K. S., Greenshaw, A., Cao, B., & Sui, J. (2022). Depression screening using a non-verbal self-association task: A machine-learning based pilot study. *Journal of Affective Disorders, 310*, 87–95. https://doi.org/10.1016/j.jad.2022.04.122

Lovibond, S.H., & Lovibond, P.F. (1995). *Manual for the Depression Anxiety Stress Scales* (2nd ed.). Sydney: Psychology Foundation.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Manual for the Profile of Mood States*. San Diego, CA: Educational and Industrial Testing Services.

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

Open Source Psychometrics Project. (2019). *Open psychology data: Raw data from online personality tests.* https://openpsychometrics.org/_rawdata/

Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science, 167*, 1258–1267. https://doi.org/10.1016/j.procs.2020.03.442

Rios, J. A., & Soland, J. (2021). Parameter estimation accuracy of the effort-moderated item response theory model under multiple assumption violations. *Educational and Psychological Measurement, 81*(3), 569–594. https://doi.org/10.1177/0013164420949896

Salthouse T. A. (2012). How general are the effects of trait anxiety and depressive symptoms on cognitive functioning? *Emotion, 12*(5), 1075–1084. https://doi.org/10.1037/a0025615

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Snyder, H. R., Kaiser, R. H., Warren, S. L., & Heller, W. (2015). Obsessive-compulsive disorder is associated with broad impairments in exec-

utive function: A meta-analysis. *Clinical Psychological Science, 3*(2), 301–330. https://doi.org/10.1177/2167702614534210

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology, 6*, 328. https://doi.org/10.3389/fpsyg.2015.00328

Srinath, K. S., Kiran, K., Pranavi, S., Amrutha, M., Shenoy, P. D., & Venugopal, K. R. (2022). *Prediction of depression, anxiety and stress levels using Dass-42* [Paper presentation]. 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Mumbai, India, 1–6.

Sun, M. K., & Alkon, D. L. (2014). Stress: Perspectives on its impact on cognition and pharmacological treatment. *Behavioural Pharmacology, 25*(5–6), 410–424. https://doi.org/10.1097/FBP.0000000000000045

Tada, M., Uchida, H., Suzuki, T., Abe, T., Pollock, B. G., & Mimura, M. (2014). Baseline difference between patients' and clinicians' rated illness severity scores and subsequent outcomes in major depressive disorder: Analysis of the sequenced treatment alternatives to relieve depression data. *Journal of Clinical Psychopharmacology, 34*(3), 297–302. https://doi.org/10.1097/JCP.0000000000000112

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*(5), 327–347. https://doi.org/10.1177/0146621609349800

Ward, M. K., & Meade, A. W. (2023). Dealing with Careless Responding in Survey Data: Prevention, Identification, and Recommended Best Practices. *Annual Review of Psychology, 74*, 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Wardenaar, K. J., Wanders, R. B. K., Roest, A. M., Meijer, R. R., & de Jonge, P. (2015). What does the Beck Depression Inventory measure in myocardial infarction patients? A psychometric approach using item response theory and person-fit. *International Journal of Methods in Psychiatric Research, 24*(2), 130–142. https://doi.org/10.1002/mpr.1467

Watson, D., & Clark, L. A. (1994). The PANAS-X: *Manual for the positive and negative affect schedule - expanded form.* University of Iowa.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education, 19*(2), 25–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L. (2020). An intelligent CAT that can deal with disengaged test taking. In H. Jiao & R. W. Lissitz (Eds.), *Application of Artificial Intelligence to Assessment* (pp. 161–174). Information Age Publishing, Inc.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012, April). *Setting response time thresholds for a CAT item pool: The normative threshold method* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Zuckerman, M., & Lubin, B. (1985). *Manual for the multiple affect adjective check list.* San Diego: Educational and Industrial Testing Service.

Figure 1

Methodological Framework for Addressing Research Question 1

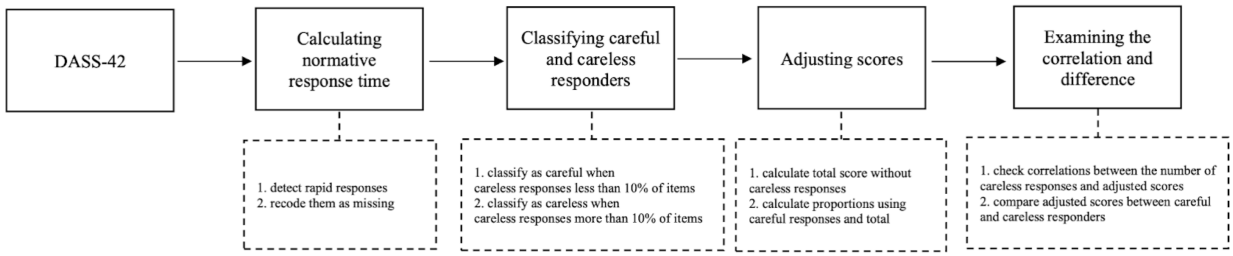


Figure 2

Methodological Framework for Addressing Research Question 2

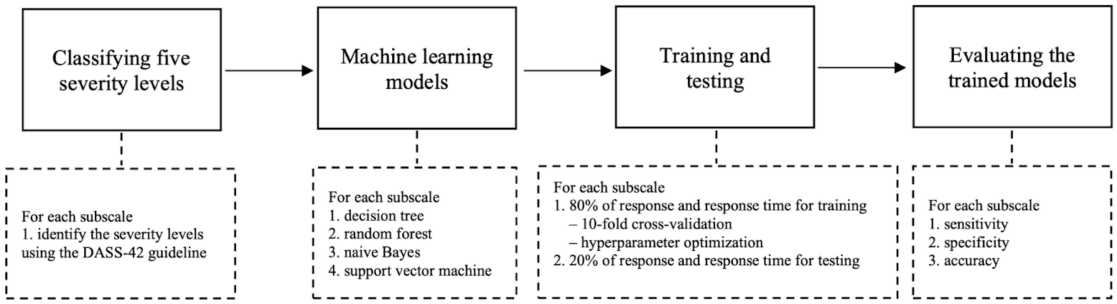


Figure 3

Comparison of DASS-42 Subscale Scores between Careful and Careless Responders

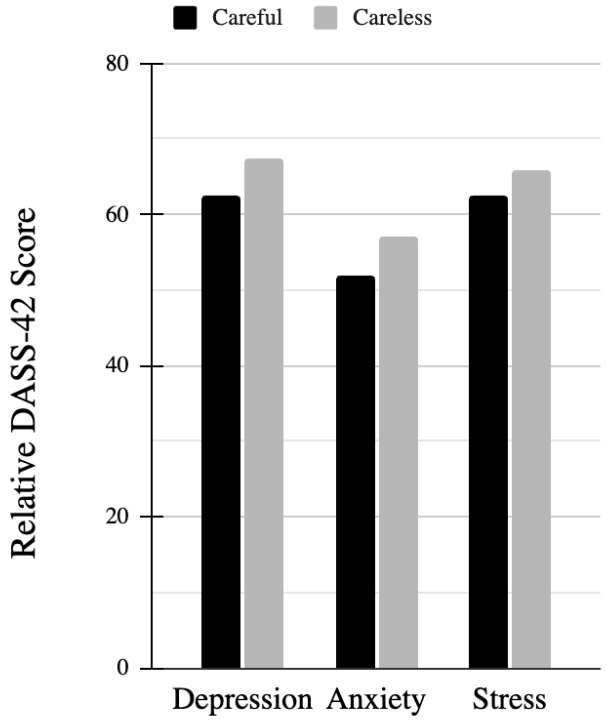
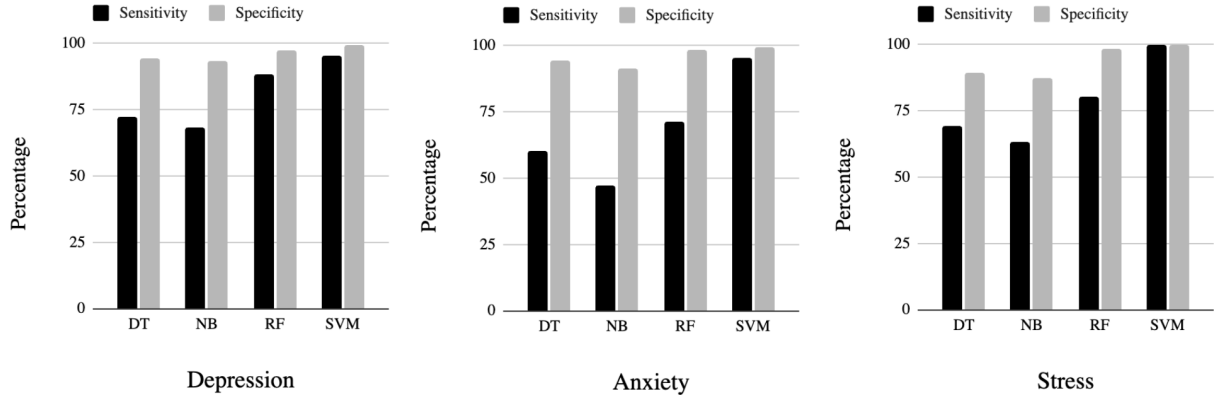


Figure 4

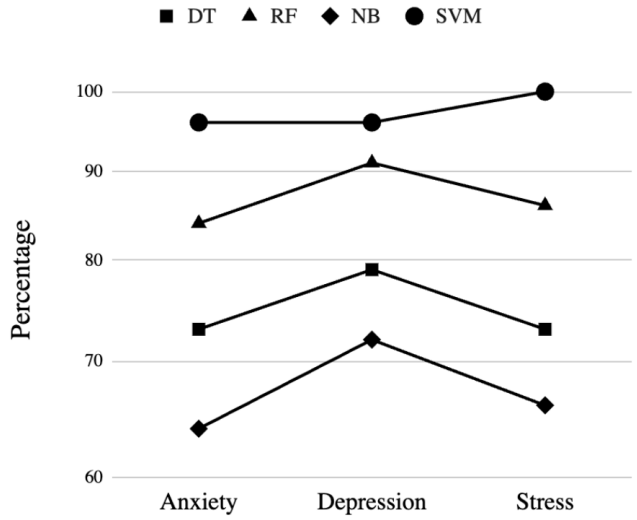
Macro-averaged Classification Metrics for Emotional Distress



Note. DT: decision tree; RF: random forest; NB: naive Bayes; SVM: support vector machine.

Figure 5

Accuracy of Classification for Machine Learning Models



Note. DT: decision tree; RF: random forest; NB: naive Bayes; SVM: support vector machine

Table 1

Sociodemographic Summary

Demographic variable	*n*	%
Education		
Less than high school	3881	10
High school	14325	38
University degree	14399	38
Graduate degree	4729	13
Urbanicity		
Rural	7892	21
Suburban	12595	33
Urban	16972	45
Gender		
Male	8365	22
Female	28864	76
Other	528	1
Ethnicity		
Asian	21910	58
Arab	311	1
Black	575	1
Indigenous Australian	23	<1
Native American	209	1
White	10236	27
Other	4562	12
Marital status		
Never married	32472	86
Currently married	4122	11
Previously married	1039	3

Table 2

Guide for Severity Levels of Emotional Distress in DASS-42

	Depression	Anxiety	Stress
Normal	0–9	0–7	0–14
Mild	10–13	8–9	15–18
Moderate	14–20	10–14	19–25
Severe	21–27	15–19	26–33
Extremely severe	28+	20+	34+

Table 3

Single Classification Metrics for Emotional Distress by Each Level

	Normal	Mild	Moderate	Severe	Extremely severe
		Depression			
DT					
Sensitivity	91.20	48.03	68.14	64.38	92.17
Specificity	96.67	95.34	93.48	93.40	95.06
RF					
Sensitivity	97.50	59.97	90.48	86.39	97.55
Specificity	97.93	98.75	96.03	97.71	**98.28**
NB					
Sensitivity	79.85	57.30	52.32	70.42	81.78
Specificity	98.28	91.69	91.33	87.64	97.50
SVM					
Sensitivity	**98.22**	**87.64**	**94.38**	**92.50**	**98.85**
Specificity	**99.37**	**99.56**	**98.68**	**99.21**	97.66

Table 3 (continued)

			Anxiety		
DT					
Sensitivity	88.55	22.84	58.27	45.64	88.42
Specificity	95.24	96.27	89.83	90.82	92.70
RF					
Sensitivity	97.73	4.80	86.62	64.28	96.60
Specificity	95.96	**99.77**	91.53	96.34	95.62
NB					
Sensitivity	92.11	0.96	52.03	12.43	83.38
Specificity	85.31	99.73	82.56	95.60	88.95
SVM					
Sensitivity	**98.22**	**87.64**	**94.38**	**92.50**	**98.85**
Specificity	**99.37**	99.56	**98.68**	**99.21**	**97.66**

Table 3 (continued)

	Normal	Mild	Moderate	Severe	Extremely severe
			Stress		
DT					
Sensitivity	88.70	41.79	65.86	69.47	80.51
Specificity	94.72	93.19	89.21	91.36	97.06
RF					
Sensitivity	96.20	48.70	89.49	90.90	87.31
Specificity	96.39	97.95	93.06	95.68	99.55
NB					
Sensitivity	79.10	45.25	54.92	62.03	78.33
Specificity	96.17	89.41	87.11	89.43	95.54
SVM					
Sensitivity	**99.96**	**100.0**	**99.82**	**100.0**	**99.91**
Specificity	**100.0**	**99.97**	**99.98**	**99.97**	**100.0**

Note. DT: decision tree; RF: random forest; NB: naive Bayes; SVM: support vector machine. Bold indicates the highest, with 70% acceptable and 80% optimal thresholds in each subscale.