scientific reports

Check for updates

OPEN

Fast, smart, and adaptive: using machine learning to optimize mental health assessment and monitor change over time

Daiana Colledani[1], Claudio Barbaranelli[1] & Pasquale Anselmi[2]✉

In mental health, accurate symptom assessment and precise measurement of patient conditions are crucial for clinical decision-making and effective treatment planning. Traditional assessment methods can be burdensome, especially for vulnerable populations, leading to decreased motivation and potentially unreliable results. Computerized Adaptive Testing (CAT) has emerged as a solution, offering efficient and personalized assessments. In particular, Machine Learning-based CAT (MT-based CATs) enables adaptive, rapid, and accurate evaluations that are more easily implementable than traditional methods. This approach bypasses typical item selection processes and the associated computational costs while avoiding the rigid assumptions of traditional CAT approaches. This study investigates the effectiveness of Machine Learning-Model Tree-based CAT (ML-MT-based CAT) in detecting changes in mental health measures collected at four time points (6-month intervals between February 2018 and December 2019). Three CATs measuring generalized anxiety, depression, and social anxiety were developed and tested on a dataset with responses from 564 participants. A cross-validation approach based on real data simulations was used. Results showed that ML-MT-based CATs produced estimates of trait levels comparable to full-length tests while reducing the number of items administered by 50% or more. In addition, ML-MT-based CATs captured changes in trait levels consistent with full-length tests, outperforming short static measures.

Keywords CAT, M5P, Machine learning, Model trees, Assessment, Mental health, Change

Mental health encompasses the diagnosis, treatment and management of various mental illnesses, disorders and conditions, as well as the promotion of well-being and personal fulfilment. In this context, accurate assessment of symptoms and objective measurement of patients' conditions are crucial. These aspects, in fact, are essential to ensure informed clinical decision making and effective treatment planning. However, the diagnosis and screening process often requires the administration of multiple tests and frequent monitoring of patients, resulting in the repetition of numerous tests and items[1,2]. While these procedures are necessary, they can be burdensome for patients, particularly the elderly or frail, who may experience boredom and fatigue in answering multiple questions that are often redundant or inappropriate for describing their condition. In this situation, motivation to answer accurately may fail and patients may provide hasty and inaccurate answers, which can lead to unreliable assessments[3–5]. Long and redundant instruments also pose limitations in the field of research, as the difficulty of administering very long scales in real mental health contexts (i.e., clinics, private practices, general healthcare contexts) reduces the possibility of conducting a systematic study of the effectiveness of treatment programs[1]. Over the years, a number of strategies have been developed to overcome these limitations. For example, brief instruments have been developed to assess the severity of many mental health conditions such as depression and anxiety (e.g., Anxiety and Depression Detector-ADD[6]; Patient Health Questionnaire-PHQ-9 [7] and PHQ-4 [8,9]; Generalized Anxiety Disorder Screener-GAD-7 [10]; Five-Item Mental Health Inventory-MHI-5 [11]), psychological distress (e.g., General Health Questionnaire-GHQ-12 Goldberg et al., 1997 [12]), mood (e.g., My Mood Monitor-M-3 Checklist[13]), and post-traumatic stress disorder (e.g., Short Posttraumatic Stress Disorder Rating Interview-SPRINT[14]; Short Forms of PTSD Checklist-PCL-C-SF[15–17])[18]. Traditionally, these instruments, as well as their full-length versions, have been developed within the framework of classical test theory[2,19–21]. Such instruments estimate the severity of a patient's condition by means of a total score that is

[1]Department of Psychology, Faculty of Medicine and Psychology, Sapienza University of Rome, Via dei Marsi 78, 00185 Rome, Italy. [2]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Piazza Capitaniato 3, 35139 Padua, Italy. ✉email: pasquale.anselmi@unipd.it

obtained by summing, or averaging, the patient's responses to a set of items that are identical for all patients. This implies that, particularly in the case of short tests, only a limited number of items are suitable for each specific respondent and can capture the different levels of severity of a disorder. Consequently, the evaluations that can be obtained may be inaccurate. Moreover, in this framework, each item, regardless of its severity, is given equal weight, so that a positive response to the item "Sometimes I feel sad" is considered equally important as a positive response to the item "Sometimes I had suicidal thoughts"[22–24].

An alternative method to improve the accuracy of mental health assessment while reducing the burden is adaptive testing[25,26]. Adaptive tests are computerized instruments (computerized adaptive test, CAT) that personalize the assessment process. CATs are designed to adaptively vary the questions administered to respondents based on their responses to previous items and the characteristics of each item. Unlike traditional tests, adaptive tests do not administer all questions to all patients, but only the questions that are most informative for each respondent at a given time. This approach enhances the efficiency and personalization of assessments, leading to more precise and meaningful measurements[27]. The majority of CATs have been developed in the context of item response theory (IRT)[21]. IRT-based CAT relies on the careful calibration of a large item pool[28,29]. The calibration considers different characteristics of each item, depending on the specific IRT model used. These characteristics include item difficulty, discrimination, pseudo-guessing, and inattention. Difficulty (parameter b) refers to the level of ability required to provide a correct response to the item or the severity of a disease necessary to provide a response indicating the presence of the disease. Discrimination (parameter a) refers to the ability of the item to effectively distinguish between individuals with high or low ability or different levels of disease severity[30]. Pseudo-guessing (parameter c) reflects the influence of chance on providing a correct response to the item or a response indicating the presence of the disease[30,31]. Inattention (parameter d) accounts for the impact of distraction on providing an incorrect response to the item or a response indicating the absence of the disease[32]. By taking these item characteristics into account, CATs present each respondent with only those items that are most informative for the current estimate of their ability level (or disorder severity level), which is progressively updated based on responses to sequentially presented items. The assessment ends when the respondent's trait level has been estimated with sufficient accuracy (variable-length CAT) or when a predetermined number of items have been administered (fixed-length CAT[33,34]). IRT-based CATs have been successfully used in several fields, including education, psychological and personality assessment, and mental health, where they have demonstrated remarkable efficiency in accurately measuring the severity of mental disorders while significantly reducing the number of administered items[35–39]. More recently, a new type of adaptive test has been proposed: the computerized adaptive diagnosis (CAD)[1,40]. These tools are not designed to provide an accurate measure of the severity of a disorder, but rather to provide accurate diagnostic categorizations (e.g., diagnosed vs. non-diagnosed; at-risk vs. not-at-risk[1]). Research has shown that CADs are valuable tools that allow for a very high degree of personalization, excellent diagnostic performance (e.g., diagnostic accuracy, sensitivity, and specificity), and great efficiency[1,22,41,42]. These tests are developed using machine learning (ML) and in particular decision tree (DT) or random forest classifier algorithms.

A more recent variant of these tools, which has not yet been widely used in diagnostic contexts, are the adaptive tests based on regression tree or model tree machine learning classifier algorithms. This type of tool uses a process similar to that of CAD. However, rather than a discrete diagnostic classification, its goal is to predict the score on a continuous variable, such as the score on a test or the severity of a disease[24]. Compared to IRT-based CATs, ML-based CATs do not require the verification of strict assumptions with respect to item dimensionality and thus are particularly useful for multifaceted constructs[43]. ML-CATs based on DT, regression tree or model tree (MT) algorithms define the selection of items to be presented to respondents according to a tree structure developed by the classifier algorithm. The tree structure is a sort of flowchart that, using a top-down approach based on a recursive divide-and-conquer process, defines a set of if-else rules[44,45]. Following the rules, it is possible to classify instances (e.g., respondents to a test) according to a specific category (e.g. diagnosis vs. non-diagnosis) or in the case of regression or model trees, it is possible to predict their score on a continuous variable (e.g., the test score[24]). The tree structure consists of a root and a series of branches and leaves. The root is the origin of the tree, namely, the node from which all branches develop. Each node is represented by a specific attribute (e.g., item/variable), and a chain of nodes from the root to a leaf (i.e., the end of a branch and the place in the tree structure where classification or score prediction takes place) forms a branch. In the first step, the algorithms select the node (i.e., the predictor) to be placed at the root of the tree and that should be subsequently divided into branches to build the tree structure. DT algorithms work by finding the best split among all the possible predictor values (e.g., in a 7-point Likert scale from 1 to 7, the splitting rule could be ≤ 2 vs. > 2) to partition the dataset into groups where the outcome values are most similar (e.g., same classification or same score). This process is carried out recursively in each derived group until the algorithm meets a stopping point (e.g., a maximum number of splits, a minimum number of cases in each group, or no prediction improvement). When the variable to be predicted is categorical, the algorithm aims to generate the "purest" nodes, that is, to divide the sample into subsets containing instances with the same classification[45]. This process is rooted in information gain[45–47] and based on the entropy of the class distribution[48] Analogously, when the variable to be predicted is continuous, the decision on which attribute is the best to split is based on minimizing the within-class variation. Specifically, the decision regarding the optimal attribute for splitting is based on the standard deviation of the class variable at each node. The standard deviation is employed as a measure of the error at each node. The algorithm calculates the expected error reduction (i.e., standard deviation) that would result from introducing a specific attribute at a specific node and selects the attribute that maximizes that reduction[45].

DTs have been demonstrated to be an effective means for instructing adaptive testing procedures. Indeed, the structure of DTs is defined by a set of nodes connected by a set of rules. In the context of a test situation, these can be represented by the test items and the specific item response options (e.g., response ≤ 2 vs. > 2), respectively. Once the structure of a DT has been defined, it can be used as a guide to administer the test items to new respondents according to an adaptive procedure. This consists simply of presenting respondents with items that are included in the branch to which they are assigned, based on the answers they progressively provide[1,22,23,40,43,49–52]. One of the key advantages of ML-based CAT over traditional IRT-based CAT is the possibility of designing the entire test in advance, based on the predefined tree structure. This allows the adaptive test to be applied to respondents immediately, bypassing the item selection process and the associated computational costs typical of IRT-based CAT[40].

Although CATs have demonstrated considerable value in clinical settings, few studies have examined their ability to capture changes in the conditions (trait levels, disorder severity levels, diagnostic categorizations) of individuals who completed the test at different time-points. This is a critical aspect for clinical practice and research, as the ability to track changes is essential for monitoring patients and verifying the efficacy of treatment programs. In the context of IRT, only a few studies have explored the efficacy of CATs in detecting change[52–54]. These studies have generally supported the notion that CATs are as effective as static tools in predicting respondents' retest scores. Furthermore, IRT-based CATs have been demonstrated to be as sensitive to change as static instruments[52–54]. However, no studies have yet examined the ability of machine learning-based CATs to detect changes. Nevertheless, investigations in this area seems to be crucial to strengthen the usefulness of these tools in clinical contexts. Should evidence confirm their effectiveness in capturing change, the implications for their application in clinical contexts would be substantial. Indeed, ML-based CATs would facilitate more efficient assessment procedures, enabling the evaluation of multiple dimensions within a single session without undue fatigue of patients. Furthermore, they would allow for more comprehensive monitoring of patient progress, thus reducing the burden on both patients and professionals. Should evidence confirm their effectiveness in capturing change, the implications for their application in clinical contexts would be substantial. Indeed, ML-based CATs would facilitate more efficient assessment procedures, enabling the evaluation of multiple dimensions within a single session without undue fatigue of patients. Furthermore, they would allow for more comprehensive monitoring of patient progress, thus reducing the burden on both patients and professionals.

The objective of this study is to assess the capacity of ML-MT-based CATs to identify fluctuations in measures of mental health. In particular, measures of anxiety, social anxiety, and depression will be considered. These mental health problems were selected because they are among the most common issues in the general population, and well-established static assessment measures are available for them[7,10,55–61]. A real data simulation approach will be employed to assess the efficiency and accuracy of ML-MT-based CATs relative to static full-length versions of the same instruments in detecting change. It is expected that the adaptive tools will provide an accurate assessment of the targeted mental health conditions while significantly reducing the number of items administered. Furthermore, these tools are expected to effectively and efficiently capture individual changes in trait levels over time, aligning with the results obtained from full-length tests.

Method

Participants

The data were retrieved from the public repository OSF (https://osf.io/jz4ge/?view_only=ea57acad6e6a4a6e86fa5a57caf493c;[55,62]). The dataset includes information from 564 Spanish university students ($M_{age} = 21.30$, $SD = 3.64$; 31.7% males) who participated in an 18-month longitudinal study assessing mental health (e.g., depression, separation anxiety, specific phobia, social anxiety, panic, agoraphobia, generalized anxiety), personality traits, and risk factors. Data were collected in four waves (T1, T2, T3, and T4) at 6-month intervals between February 2018 and December 2019. Specifically, the dataset includes data from participants who completed mental health assessments at least at one time-point (i.e., 564 at T1, 359 at T2, 289 at T3, and 276 at T4). Participants received financial compensation for their participation at each assessment wave (€5, €10, €15, and €15 at T1, T2, T3, and T4, respectively). Further details in Table 1. All participants provided informed consent before participating in the study[55,62]. The experimental protocol from which the data used in the present work derive was approved by the Ethics Committee of the Universitat Jaume I (Castelló de la Plana, Castellón, Spain)[55]. All methods were carried out in accordance with relevant guidelines and regulations.

Measures

In this work three mental health measures were considered: general anxiety, depression, and social anxiety.

In particular, the Spanish versions of two DSM-5 anxiety severity measures, namely Generalized Anxiety Disorder Dimensional Scale (GAD-D) and Social Anxiety Disorder Dimensional Scale (SAD-D), were used to assess generalized anxiety and social anxiety[56,62]. Each scale consists of 10 items rated on a 5-point scale (from 0 = never to 4 = all of the time). All items require participants to rate how often they have experienced the feelings described in the items over the past six months (sample item for generalized anxiety: "During the past six months, I have felt anxious, worried, or nervous"; sample item for social anxiety: "During the past six months,

	T1	T2	T3	T4
N	564	359	289	276
Percentage of males	31.70	28.40	28.00	26.10
Percentage of females	68.30	71.60	72.00	73.90
Mean age	21.30	21.70	22.10	22.40
SD age	3.64	3.31	3.59	3.30

Table 1. Descriptive statistics for respondents at the different time points.

I have felt moments of sudden terror, fear or fright in social situations"). For both scales, the higher the score, the greater the severity of anxiety. Validation studies in the Spanish context supported good reliability for both scales (Cronbach's alpha from 0.92 to 0.93 and from 0.90 to 0.92; and McDonald's Omega from 0.92 to 0.94 and from 0.91 to 0.92 for SAD-D and GAD-D, respectively), as well as good structural, convergent, and discriminant validity[55,62]. Scalar longitudinal invariance for both measures was also established for both measures[62]. Cut-off scores of ≥ 13 (sensitivity = 0.88 and specificity = 0.65) and ≥ 14 (sensitivity = 0.88 and specificity = 0.75) were proposed for GAD-D and SAD-D, respectively[63].

Depression was measured using the Patient Health Questionnaire-9 (PHQ-9)[7,10], which consists of 9 items scored on a 4-point response scale (from 0 = never to 3 = everyday) and assesses the diagnostic criteria for major depression in the DSM-IV. Higher scores indicate greater depression. However, whereas the original PHQ-9 focuses on the past 15 days, the version considered in Vidal-Arenas et al.[62] focuses on the past 6 months. The PHQ-9 is one of the most widely used instruments for screening and diagnosing major depressive disorder[64]. The literature suggests that it has good diagnostic accuracy, validity, and reliability[64]. Typically, a score ≥ 10 is suggested as the optimal cut-off point[7,65,66] for detecting depression, with sensitivity and specificity ranging from 0.37 to 0.98 and 0.42 to 0.99, respectively[64]. A shorter version of the instrument, known as the PHQ-2, is also available[67]. It consists of the first two items of the PHQ-9, which refer to depressed mood and anhedonia, the two core symptoms of depression. The validation study of the PHQ-2 indicated a cut-off score ≥ 3 as the optimal threshold for identifying depression, with a sensitivity of 0.83 and a specificity of 0.92[67].

Analyses

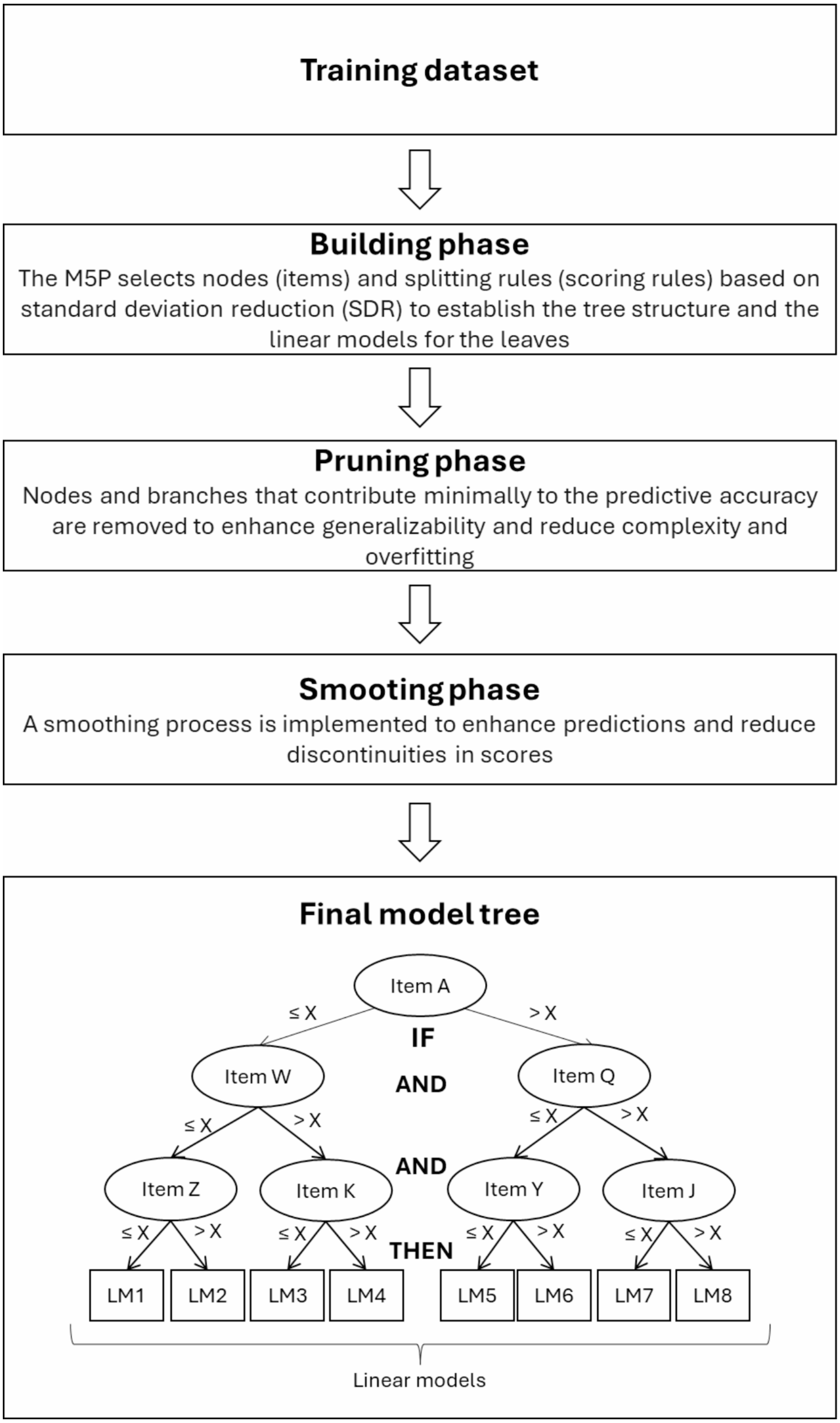
To evaluate the ability of ML-MT-based CATs to identify fluctuations in measures of mental health, a cross-validation approach based on real data simulation was used. For each of the three mental health scales (GAD-D, PHQ-9, and GAS-D), two datasets were created: a training dataset to train the ML-MT algorithm and a testing dataset to evaluate the performance of the algorithm on unseen data[68].

In particular, for the GAD-D scale, the testing dataset was constructed by selecting all respondents who completed the scale at both T1 and T4 (N = 274, M_{age} at T1 = 20.92, SD = 3.37; Females 73.4%), while the training dataset included the responses to the GAD-D scale at all time-points for the remaining participants (the dataset includes 431 response patterns from 290 respondents; M_{age} at T1 = 21.65, SD = 3.84; Females 63.4%).

The training dataset was used to train the M5P algorithm using the software Weka 3.8.5 (Waikato Environment for Knowledge Analysis;[68]). The M5P is an ML-MT algorithm that deals with continuous rather than discrete outcome variables[69]. It is the expanded version of M5 algorithm that was originally developed by Quinlan[70]. The algorithm aims to identify a classification function to predict the value of a target variable based on a set of input variables. In this case, the algorithm was used to predict the GAD-D score as accurately as possible based on responses to a limited number of items. The algorithm determines which elements to place at the different levels of the DT (both the root node and subsequent nodes), as well as the splitting rules (i.e., item scoring rules) that define its branches and the tree structure. During the tree development phase (building phase), the algorithm progressively selects items and splitting rules based on the standard deviation (*SD*) of the values of the outcome variable (GAD-D sum score) within the different segments of the data set that are progressively created. The *SD* serves as an error indicator at each node, guiding the algorithm to select the attribute that maximizes error reduction (i.e., standard deviation reduction; *SDR*) at that particular node[71]. After building the tree, the algorithm developed a linear regression model for each leaf using the data associated with its specific branch. Then, pruning was carried out to address overfitting issues and obtain the final model[45]. A smoothing process was also implemented to enhance predictions and reduce discontinuities in scores[45,72]. A visual representation of the process is depicted in Fig. 1.

After the tree structure was developed using the training dataset, it was used to simulate an adaptive assessment on the testing dataset. The algorithm performance was evaluated in terms of (a) accuracy of the ML-MT-based CAT in reproducing the score obtained with the full-length GAD-D, (b) capability of the ML-MT-based CAT in reproducing the same change in generalized anxiety at different time points as that obtained with the full-length GAD-D, and (c) efficiency of the ML-MT-based CAT compared with the full-length GAD-D. The accuracy in reproducing the score obtained with the full-length GAD-D was evaluated using three common indices: correlation, mean absolute error (MAE), and *t*-tests. Correlation evaluates the extent to which the test score obtained using the CAT are associated to the scores obtained by administering all the scale items (i.e., the actual score). High correlations indicate good performance. MAE estimates the average (in absolute terms) difference (error) between CAT and full-length test scores (i.e., predicted and actual outcome values). A small MAE indicates that the CAT scores closely match the actual scores. The *t*-tests assess the null hypothesis that the scores on the ML-MT-based CAT do not differ from those obtained by administering the full-length GAD-D.

The capability in reproducing the same change in generalized anxiety as that obtained with the full-length GAD-D was evaluated using three methods. In particular, *t*-tests were used to determine whether any differences between T1 and T4 that were detected using the full-length scale scores could also be detected using the ML-MT-based CAT scores. In addition, the significance of the difference in GAD-D scores between T1 and T4 was calculated for each individual using the standard error of measurement $SEM = SD \times \sqrt{1-\alpha}$ (*SD* and Cronbach's α were calculated on the training data set). This calculation was carried out for both the full-length test scores and the ML-MT-based CAT scores. Then, the number of cases in which the full-length GAD-D and the CAT reached the same conclusion regarding a significant change was determined. Finally, the cut-off score recommended in the literature for the GAD-D (≥ 13 [63]) was used to categorize individuals as above or below the threshold. This cut-off score was used to categorize respondents at both T1 and T4, using both the full-length test and the CAT scores. The degree of agreement between these two methods in detecting changes in classification between the two time-points was then calculated.



Training dataset

Building phase
The M5P selects nodes (items) and splitting rules (scoring rules) based on standard deviation reduction (SDR) to establish the tree structure and the linear models for the leaves

Pruning phase
Nodes and branches that contribute minimally to the predictive accuracy are removed to enhance generalizability and reduce complexity and overfitting

Smooting phase
A smoothing process is implemented to enhance predictions and reduce discontinuities in scores

Fig. 1. M5P algorithm process flowchart.

To explore potential patterns in the disagreement between methods in detecting changes in scores or diagnostic classifications, *t*-tests, χ^2 tests, and point-biserial correlations ($r_{\mathrm{p.bis}}$) were conducted. Specifically, *t*-tests and χ^2 tests were used to investigate whether age or gender influenced the performance of CATs in terms of agreement with the full-length test (change in scores and change in diagnostic categorization), while point-biserial correlations were used to investigate whether the level of agreement was associated with the number of items presented in the CATs.

A similar procedure was used also for the depression scale. Two datasets were constructed. Specifically, the testing dataset was built by selecting all respondents who completed the PHQ-9 at both T1 and T4 ($N = 276$, M_{age} at T1 = 20.91, *SD* = 3.37; Females 73.6%), while the training dataset included the responses to the PHQ-9 scale at all time-points for the remaining participants (the dataset includes 427 response patterns from 288 respondents; M_{age} at T1 = 21.67, *SD* = 3.85; Females 63.2%). A further comparison was conducted on the testing dataset in addition to those carried out in GAD-D. Specifically, the ability of the ML-MT-based CAT to detect change was compared not only with that of the full-length test but also with that of the 2-item static short version of the instrument (i.e., the PHQ-2). The comparisons were performed both considering the test scores (using the *SEM* computed in the training data set) and the classifications made using the cut-off score recommended in the literature (i.e., ≥ 3)[67]. To compare the performance of CAT and the static short form in terms of agreement with the full-length version, McNemar's tests were conducted.

A methodology similar to that used for generalized anxiety was also used for social anxiety. In this case, the testing dataset included all subjects who responded to all waves ($N = 237$, M_{age} at T1 = 20.86, *SD* = 3.00; Females 73.8%), while the training dataset included the measurements at all the available time-points from all the remaining respondents (the dataset includes $N = 539$ response patterns from 327 respondents; M_{age} at T1 = 21.61, *SD* = 4.01; Females 64.2%). However, comparisons were made across all the four time points. This approach aimed to provide a more nuanced exploration of the capability of the ML-MT-based CAT to detect changes in mental health measures over time, extending the analysis beyond two-time points to include more waves.

Results

Attrition was not directly addressed in the primary analyses and all available data were used. However, age and gender differences were assessed between participants who completed all waves and those with missing data in one or more waves. Although statistically significant differences were observed for both gender ($\chi^2(1,564) = 5.87$, $p = .02$, Cramér's $V = 0.10$) and age ($t(560) = 2.41$, $p = .02$, Cohen's $D = 0.20$), the effect sizes were negligible to small.

Power analyses were performed to determine whether the sample sizes of the datasets were sufficient to detect small effect sizes in the *t*-tests and correlation analyses, using a significance level (α) of 0.05 (assuming two-tailed tests) and a power ($1 - \beta$) of 0.80. For *t*-tests, a total sample size of at least $N = 199$ was required to detect a small effect ($d = 0.2$). Similarly, for the correlational analyses, a sample size of $N = 193$ was required to detect a small effect ($r = .2$; $N = 191$ was required for point-biserial correlations). In our analyses with the GAD-D, PHQ-9, and SAD-D, the sample sizes consistently exceeded these requirements, with $N = 274$, 276, and 237, respectively.

Generalized anxiety (GAD-D)

The M5P algorithm applied to the responses to the 10 items of the GAD-D scale in the training dataset resulted in a tree structure that placed item 2 at the root node. The tree contained 29 leaves, with branches consisting of 3 to 8 nodes. Considering that some items may appear at multiple levels within the tree (albeit with different splitting/scoring rules), the unique items appearing in the different branches ranged from 2 to 7. It is interesting to note that the algorithm placed item 2 (i.e., "…felt anxious, worried, or nervous"), which refers to particularly salient content for the assessment of generalized anxiety, at the root of the DT.

The tree structure was used to simulate the administration of a CAT procedure on the response patterns of the testing dataset. The performance of the adaptive procedure yielded favorable results. The CAT procedure allowed for the estimation of test scores by administering an average of 4.49 items at T1 and 4.39 items at T4, resulting in item savings of more than 50% (i.e., from 55.1 to 56.1%).

The correlation between the full-length test scores (i.e., obtained on all 10 items of the GAD-D) and the CAT scores was 0.89 at T1 and 0.88 at T4 (Fig. 2). The MAE was 1.62 at T1 and 1.56 at T4. The *t*-tests performed to examine the comparability of the scores obtained on the adaptive test with those obtained by administering the full-length scale showed that the CAT scores did not differ from the full-length scores both at T1 and T4 (Table 2). The correlation between the test scores at T1 and T4 computed based on the full-length scores was 0.42, while it was 0.36 when computed based on CAT scores. According with the Steiger[73] test the two coefficients were not significantly different ($z = 1.53$, $p = .13$).

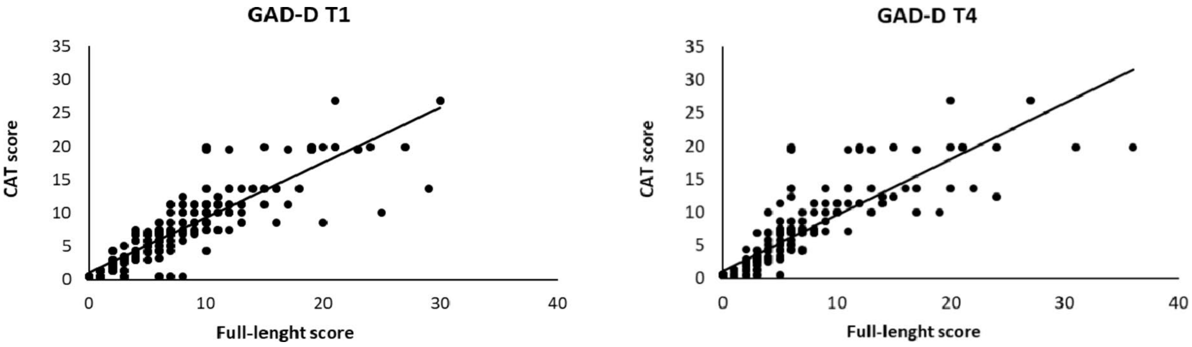


Fig. 2. Scatterplots of full-length test scores and CAT scores on the GAD-D at waves T1 and T4. *Note*. $r = 0.89$ and 0.88, $ps \le .001$, for T1 and T4, respectively.

	Full-length test scores (10 items)		CAT scores		
	Mean	*SD*	Mean	*SD*	*t*
T1	6.98	5.71	6.78	5.33	1.21; *d* = 0.07
T4	5.35	5.80	5.64	5.59	− 1.71; *d* = − 0.10
t	4.35***; *d* = 0.26		3.08**; *d* = 0.19		

Table 2. *t*-tests comparing full-length test scores and CAT scores on the GAD-D at waves T1 and T4. *Note.* *d* = Cohen's *d* effect size measure. For all *t*-tests, *df* = 274. * $p < .05$. ** $p < .01$. *** $p \leq .001$.

A significant reduction in GAD-D scores was observed from T1 to T4 for both the full-length and CAT scores (Table 2). Using the *SEM* (*SEM* = 1.92; *SD* on the training sample = 6.26, Cronbach's alpha = 0.906) a significant ($p < .05$) difference between the GAD-D scores at T1 and T4 was observed in 88 respondents when considering the full-length test scores and in 97 respondents when considering the CAT scores. Overall, the two testing methods reached the same results in 219 out of 274 cases, with an agreement of 79.9%.

When applying the cut-off score recommended in the literature for the GAD-D (≥ 13)[63], 33 respondents were above the threshold at T1 using the full-length test scores, while 28 were above the threshold using the CAT scores. At T4, 30 respondents were above the threshold using the full-length test scores, while 29 were above the threshold using the CAT scores. The agreement in detecting changes in categorization between the two methods was 87.6% (240 cases out of 274).

The CAT exhibited comparable performance in detecting changes in agreement with the full-length test for both male and female participants, with no observable age-related differences. This was observed in changes in categorical classification (gender: $\chi^2(274) = 2.67$, $p = .10$; age: $t(272) = 0.47$, $p = .64$) and in changes in scores (gender: $\chi^2(274) = 0.21$, $p = .64$; age: $t(272) = -1.23$, $p = .22$). Moreover, the agreement between the CAT and full-length test scores in detecting changes in scores from T1 to T4 was not significantly correlated with the number of items presented in the CATs ($r_{\text{p.bis}} = -0.029$ and 0.08, *ps* > .05, for T1 and T4, respectively). Regarding the agreement between the two methods in detecting changes in diagnostic categorizations, significant correlations between the number of items presented in the CATs at T1 and T4 were observed, but these correlations were small in size ($r_{\text{p.bis}} = 0.12$ and 0.21, *ps* < .05 and < .01, for T1 and T4, respectively).

Depression (PHQ-9)

Applying the M5P algorithm to the responses to the PHQ-9 items in the training dataset produced a tree structure with 26 leaves. The algorithm placed item 4 at the root node and developed branches containing 3 to 8 nodes (3 to 8 items). Interestingly, the item placed at the root of the DT asks about lack of energy (i.e. "…feeling tired or having little energy"), which is a crucial feature for the diagnosis of depression[7].

The tree structure was used to simulate the administration of a CAT on the response patterns of the testing dataset, and the performance of the CAT was compared with that of the full-length scale. The CAT procedure showed great efficiency allowing to complete assessments by administering an average of 4.46 items at T1 and 4.39 items at T4, and resulting in item savings of approximately 50% (i.e., from 55.4 to 56.1%).

The results showed a correlation between the full-length test scores (i.e., the scores obtained on all 9 items of the PHQ-9) and the CAT scores of 0.88 at T1 and 0.91 at T4 (Fig. 3). The MAE was 1.34 at T1 and 1.16 at T4. Moreover, the *t*-tests performed to examine the comparability of the scores obtained on the CAT with those obtained by administering the full-length scale showed that the CAT scores did not differ from the full-length test scores both at T1 and T4 (Table 3).

A significant decrease in PHQ-9 scores was observed from T1 to T4 for both the full-length test scores and the CAT scores (Table 3). The correlation between the test scores at T1 and T4 computed based on the full-length scores was 0.49, while it was 0.43 when computed based on CAT scores. According with the Steiger test[73] the two coefficients were significantly different ($z = 1.99$, $p = .15$) but the effect size was negligible (Cohen's *q* of 0.08)[74,75].

Using the *SEM* (*SEM* = 1.66; *SD* on the training sample = 3.80, Cronbach's alpha = 0.810), a significant ($p < .05$) difference between the PHQ-9 scores at T1 and T4 was observed in 154 respondents when considering the full-length test scores and in 144 respondents when considering the CAT scores. Overall, the two testing methods reached the same results in 208 out of 276 cases, with an agreement in detecting score changes of 75.4%.

Applying the cut-off score recommended in the literature for the PHQ-9 (≥ 10)[7], 48 respondents were above the threshold at T1 using the full-length test scores, whereas 35 were above the threshold using the CAT scores. At T4, 48 respondents were above the threshold using the full-length test scores, while 29 were above the threshold using the CAT scores. The agreement in detecting changes in categorization between the two methods was 87.7% (242 out of 276 cases).

According to χ^2 and *t*-test analyses, the CAT yielded comparable results in detecting changes in alignment with the full-length test across both males and females, with no significant age-related differences. This finding was consistent for changes in diagnostic classification (gender: $\chi^2(276) = 0.659$, $p = .40$; age: $t(274) = -1.24$, $p = .216$) and changes in scores (gender: $\chi^2(276) = 0.097$, $p = .755$; age: $t(274) = -0.066$, $p = .947$). Moreover, the agreement between the CAT and the full-length test in detecting score changes from T1 to T4 was not significantly associated with the number of items administered in the CATs ($r_{\text{p.bis}} = -0.037$ and -0.073, *ps* > .05, for T1 and T4, respectively). Regarding the correlations between the agreement of the two methods in detecting changes in diagnostic categorization and the number of items presented in the CATs, a non-significant correlation was

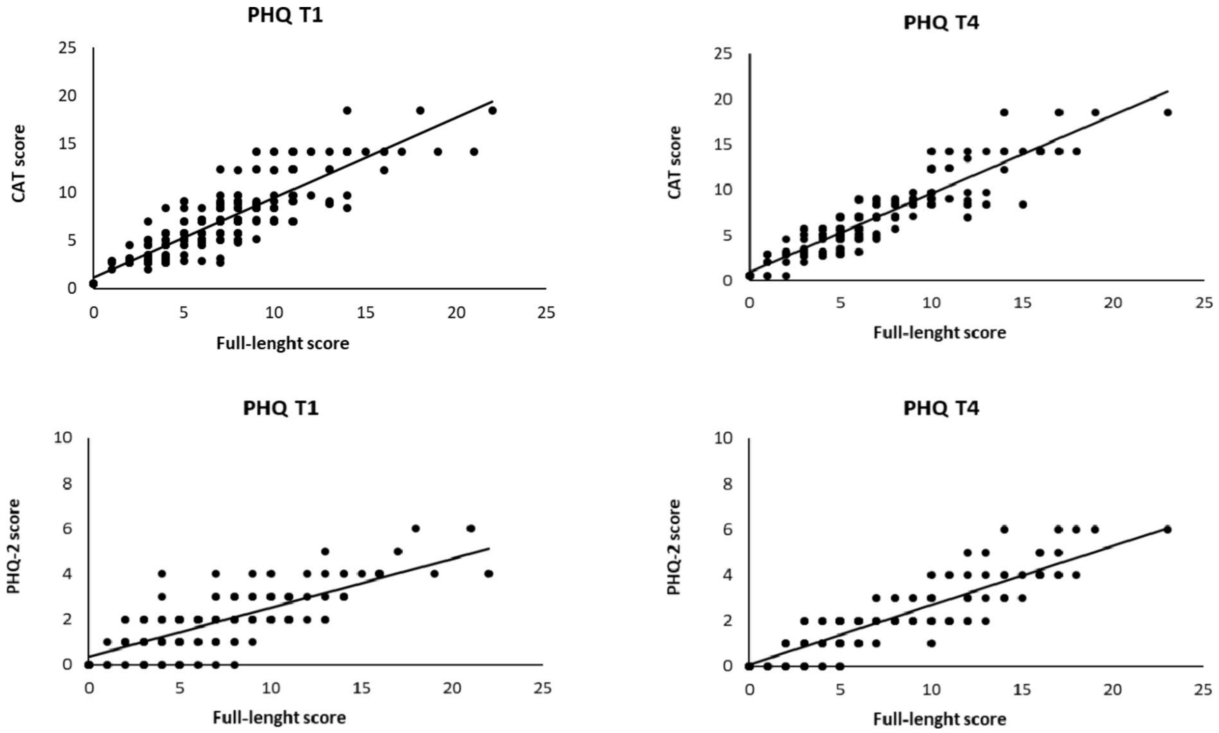


Fig. 3. Scatterplots of full-length test scores, short scale scores (PHQ-2), and CAT scores on the PHQ-9 at waves T1 and T4. *Note*. For the correlations between full-length test scores and CAT scores, $rs = 0.88$ and 0.93, $ps \leq .001$, for T1 and T4, respectively. For the correlations between full-length test scores and PHQ-2 scores, $rs = 0.77$ and 0.88, $ps \leq .001$, for T1 and T4, respectively.

	Full-length test scores (9 items)		CAT scores		
	Mean	*SD*	Mean	*SD*	*t*
T1	6.67	3.80	6.70	3.59	– 0.30; $d = -0.018$
T4	5.85	4.38	6.01	4.07	– 1.64; $d = -0.10$
t	3.25***; $d = 0.20$		2.78**; $d = 0.17$		

Table 3. *t*-tests comparing full-length test scores and CAT scores on the PHQ-9 at waves T1 and T4. *Note.* d = Cohen's d effect size measure. For all *t*-tests, $df = 275$. * $p < .05$. ** $p < .01$. *** $p \leq .001$.

found at T1 ($r_{\text{p.bis}} = 0.035$, $p > .05$), while a significant but small correlation was observed at T4 ($r_{\text{p.bis}} = -0.152$, $p < .05$).

The first two items of the PHQ-9 can be used to compute a short measure of depression called PHQ-2. When the correlation between the test scores at T1 and T4 was computed based on the full-length version of the test, it was 0.49, while it was 0.35 when computed using the short version of the instrument. According to the Steiger test, the two coefficients were significantly different ($z = 3.22$, $p \leq .001$) with a small effect size (Cohen's $q = 0.08$)[74,75]. Relying on the scores from the short static version of the scale, a significant difference between scores at T1 and T4 was observed for 73 out of 276 respondents, with 61.96% agreement with the full-length version (171 out of 276 cases). McNemar's test indicated that this performance was significantly poorer compared with that of the CAT ($\chi^2(1) = 12.3$, $p \leq .001$).

The literature suggests a cut-off score ≥ 3 as the optimal threshold for identifying major depression[67]. Using this cut-off in the PHQ-2, 47 people were above the threshold at T1, while 39 were above the threshold at T4. The agreement between the full-length measure (i.e., the PHQ-9) and the short static measure (i.e., the PHQ-2) in detecting a change in classification was 81.16% (224 cases out of 276). Also in this case, McNemar's test indicated that this performance was significantly worse than that of the CAT ($\chi^2(1) = 5.59$, $p < .05$).

Social anxiety (SAD-D)

When applying the M5P algorithm to the responses to the 10 items of the SAD-D in the training dataset, a tree structure with 26 leaves was generated. The algorithm developed branches containing 4 to 7 nodes (3 to 7 items). Interestingly, item 1, which focuses on discomfort in social situations (i.e., "…felt moments of sudden terror, fear, or fright in social situations") and represents a key feature of Social Anxiety Disorder (SAD), was placed at the root node by the algorithm.

The tree structure was used to instruct the CAT procedure that was applied to the response patterns of the testing dataset. The CAT procedure showed great efficiency, allowing assessments to be completed with an average of 4.59, 4.43, 4.44, and 4.35 at T1, T2, T3, and T4, respectively, resulting in item savings of more than 50% (i.e., from 54.1 to 56.5%).

The results showed a correlation between the full-length test scores (i.e., all 10 items of the SAD-D) and the CAT scores of 0.90, 0.94, 0.92, and 0.94 at T1, T2, T3, and T4, respectively (Fig. 4). The correlations between the SAD-D scores at different time points (i.e., T1-T2, T1-T3, T1-T4, T2-T3, T2-T4, T3-T4) ranged from 0.46 (between T1 and T4) to 0.65 (between T2 and T3; *ps* < .001) using the full-length scale and from 0.38 (between T1 and T4) to 0.60 (between T2 and T3; *ps* < .001) using the CAT. The Steiger[73] test showed that there was no significant difference in the correlations between T1 and T2 ($z = 1.95$, $p = .05$), T1 and T3 ($z = 1.48$, $p = .14$), and T2 and T4 ($z = 0.69$, $p = .49$) when calculated using the full-length test or the CAT scores. In the remaining cases, the differences between the correlations were significant: T1-T4 ($z = 2.61$, $p = .01$), T2-T3 ($z = 2.23$, $p = .03$), and T3-T4 ($z = 3.78$, $p \leq .001$). However, the effect sizes were negligible to small, with Cohen's *q* of 0.093, 0.092, and 0.15 for T1-T4, T2-T3, and T3-T4, respectively[74,75].

The MAE was 1.68, 1.28, 1.33, and 1.19 at T1, T2, T3, and T4, respectively. The *t*-tests performed to examine the comparability of the scores obtained with CAT and those obtained by administering the full-length scale showed that the CAT scores did not differ significantly from the full-length scores at T2, T3, and T4, while a significant difference was observed at T1, although the effect size was negligible (Cohen's $d = -0.15$; Table 3).

A significant decrease in SAD-D scores was detected at all waves (i.e., T1-T2, T1-T3, T1-T4, T2-T3, T2-T4) except between T3 and T4. This finding was observed for both the full-length test scores and the CAT scores (Table 4). Using the *SEM* (*SEM* = 1.81; *SD* on the training sample = 6.40, Cronbach's alpha = 0.920), a significant ($p < .05$) difference between the SAD-D scores across the four waves was observed for an average of 59 respondents using the full-length test and 56 respondents using the CAT (see Table 5). On average, the two testing methods yielded the same results in 209 out of 297 cases, resulting in an agreement rate of 88.19%. When applying the cut-off score recommended in the literature for the SAD-D (≥ 14;[63]), on average, 23 respondents were identified as meeting the categorization criteria across waves using the full-length test, and 20 using the CAT (Table 5). On average, the two testing methods yielded the same results in 222 out of 297 cases, with an agreement rate of 93.7%.

The *t*-tests and χ^2 tests used to assess potential age- and gender-related differences in CAT performance for detecting changes (in test scores or diagnostic categorization) showed no statistically significant differences for most comparisons (i.e., T1-T2, T1-T3, T1-T4, T2-T3, and T3-T4). A significant gender difference in change detection for diagnostic categorization was found only between T1 and T3, although the effect size was small (χ^2 = 5.73, $p = .017$, Cramér's $V = 0.16$). Furthermore, the agreement between CAT and full-length test scores in

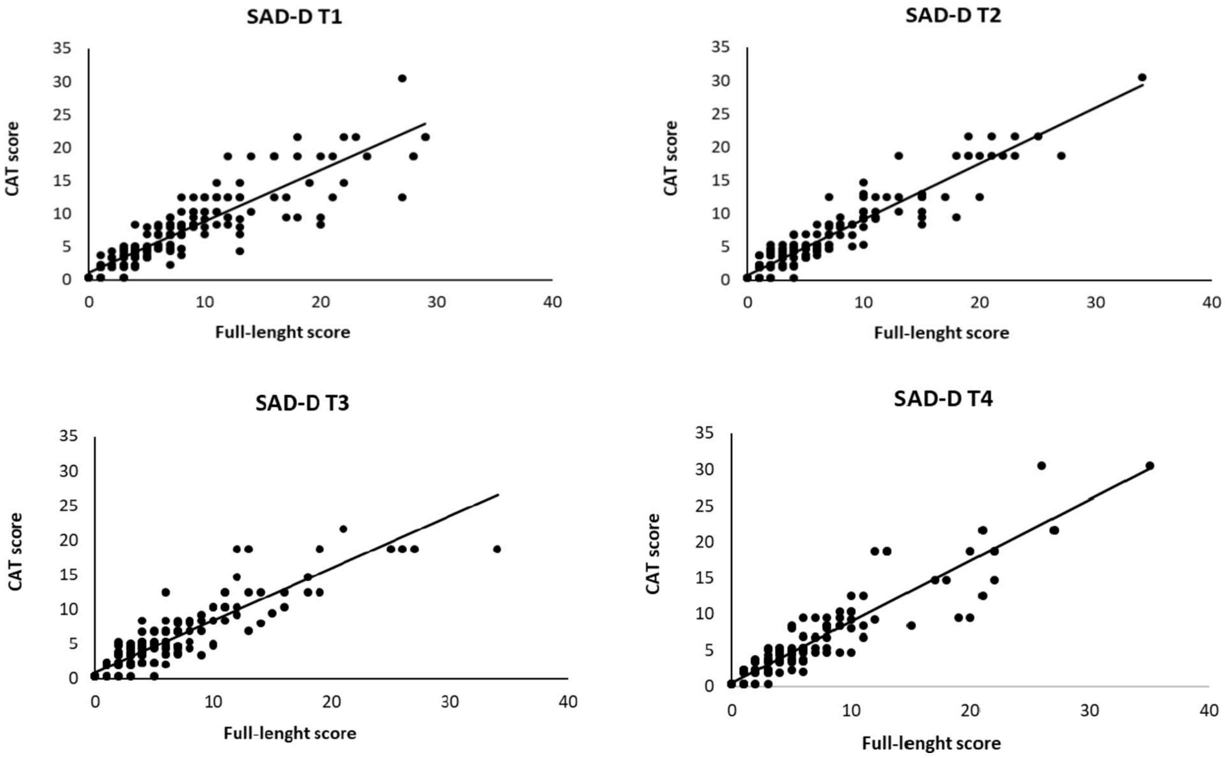


Fig. 4. Scatterplots of full-length test scores and CAT scores on the SAD-D at waves T1, T2, T3, and T4. *Note*. $r = 0.91$, 0.95, 0.92, and 0.94, $ps \leq .001$, for T1, T2, T3, and T4, respectively.

	Full-length test scores (10 items)		CAT scores		
	Mean	*SD*	Mean	*SD*	*t*
T1	6.82	6.35	6.42	5.46	2.28*; $d = 0.15$
T2	5.15	5.87	5.02	5.22	1.01; $d = 0.07$
t	4.76***; $d = 0.31$		4.27***; $d = 0.28$		
T1	6.82	6.35	6.42	5.46	
T3	4.40	5.47	4.17	4.50	1.59; $d = 0.10$
t	6.36***; $d = 0.41$		6.69***; $d = 0.43$		
T1	6.82	6.35	6.42	5.46	
T4	4.02	5.87	3.89	5.27	0.95; $d = 0.06$
t	6.77***; $d = 0.44$		6.51***; $d = 0.42$		
T2	5.02	5.22	5.15	5.87	
T3	4.17	4.50	4.40	5.47	
t	3.00**; $d = 0.20$		2.44*; $d = 0.16$		
T2	5.02	5.22	5.15	5.87	
T4	3.89	5.27	4.02	5.87	
t	3.46***; $d = 0.22$		3.15**; $d = 0.20$		
T3	4.17	4.50	4.40	5.47	
T4	3.89	5.27	4.02	5.87	
t	0.88; $d = 0.06$		1.19; $d = 0.08$		

Table 4. *t*-tests comparing full-length test scores and CAT scores on the SAD-D at waves T1, T2, T3, and T4. *Note.* d = Cohen's d effect size measure. For all t-tests, $df = 236$. * $p < .05$. ** $p < .01$. *** $p \leq .001$.

	T1-T2	T1-T3	T1-T4	T2-T3	T2-T4	T3-T4	Average
Score change with full-length test	60	79	75	44	50	47	59
Score change with CAT	59	71	74	39	48	43	56
Agreement btw full-length test and CAT	208 (87.76%)	201 (84.81%)	210 (88.61%)	212 (89.45%)	209 (88.19%)	213 (89.87%)	209 (88.19%)
Categorization change with full-length test	19	27	28	22	27	13	23
Categorization change with CAT	20	23	26	15	22	11	20
Agreement btw full-length test and CAT	224 (94.51%)	219 (92.41%)	223 (94.09%)	216 (91.14%)	224 (94.51%)	227 (95.78%)	222 (93.67%)

Table 5. Score change, categorization change and agreement in detecting change across the four waves using the full-length scales and the CAT.

detecting score changes across most wave pairs was not significantly related to the number of items administered in the CATs ($r_{\text{p.bis}}$ ranging from −0.084 to 0.116, *p*s > .05). Significant correlations were observed only for the detection of change between T2 and T4, and between T1 and T3, with the number of items in CATs at T4 and T1, respectively ($r_{\text{p.bis}} = -0.211$, $p < .001$; $r_{\text{p.bis}} = -0.149$, $p < .05$). Similarly, the agreement between CAT and full-length test scores in detecting changes in diagnostic categorizations across most wave pairs was not significantly associated with the number of items in the CATs ($r_{\text{p.bis}}$ ranging from −0.116 to 0.065, *p*s > .05). A small but significant correlation was found only for change detection between T2 and T4 with the number of items in the CAT at T4 ($r_{\text{p.bis}} = -0.162$, $p < .05$).

In essence, the findings of the study revealed that CATs based on ML-MT yielded trait-level estimates comparable to those of full-length tests, while reducing the number of items by over 50% (from 54.1 to 56.5%). Furthermore, the ML-MT-based CATs demonstrated the ability to detect changes in trait levels and diagnostic categorization over time, showing substantial agreement with the full-length tests (agreement for change in score ranged from 75.4 to 89.9%; agreement for change in diagnostic categorization ranged from 87.6 to 95.8%). Additionally, in the measure of depression, the CAT demonstrated significantly superior accuracy compared to static short forms (the CAT exhibited 87.7% agreement in change in categorization, compared to 81.2% for the static short form, and 75.4% agreement in score change compared to 62.0% for the static short form).

Discussion

The aim of this study was to investigate whether CATs based on a ML-MT algorithm (namely, the M5P algorithm) are effective in detecting changes in mental health measurements collected in different time-points. CATs are highly valued and convenient instruments for psychological assessment because they reduce the burden of testing while ensuring high assessment accuracy[25,26]. This, in turn, promotes the possibility of conducting multiple assessments over time, even with a larger number of variables. CATs are generally developed within the framework of IRT[21]. Over the years, IRT-based CATs have demonstrated their efficiency and accuracy in estimating respondents' trait levels, and some evidence has been gathered regarding their effectiveness in detecting fluctuations in trait levels in measures obtained at different time points[52–54].

In recent times, ML has emerged as a powerful tool in the field of psychological assessment. Research has demonstrated that it can be a valuable resource for the development of computerized adaptive diagnostic tests (also known as computerized adaptive diagnostics, or CADs)[1,40]. To date, results obtained in this field suggest that CADs are efficient and accurate tools for conducting assessments aimed at categorizing individuals as having or not having a certain disease[1,22,41]. ML-MT based CATs are an evolution of CAD. Although these tools work in a similar way to CAD, they offer a wider range of possibilities. ML-MT-based CATs allow not only the categorization of individuals, but also the estimation of the severity of a disorder (or the estimation of trait levels). This feature is particularly important in mental health, as it not only facilitates the identification of individuals at risk or with specific conditions for screening purposes, but also enables a detailed monitoring of their conditions to guide therapeutic interventions. ML-MT-based CATs also offer an alternative to traditional IRT-based CATs. In fact, they are more suitable for assessing multidimensional constructs and require less effort to implement the computer administration protocol compared to IRT-based CATs[40,43].

While ML-based CATs have significant potential in the field of mental health assessment, no study has yet investigated their ability to accurately assess mental health variables to detect changes in measurements over time. The aim of this study was to fill this gap. The results showed that, overall, ML-MT-based CATs allow for the assessment of psychological variables with an accuracy comparable to that of full-length versions, while achieving item savings of over 50% (from 54.1 to 56.5%). Indeed, the results indicated that the scores from ML-MT-based CATs and the full-length tests did not differ significantly. Furthermore, the MAEs were consistently small and below the SEM values of the tests. This suggests that ML-MT-based CATs do not introduce biases in the assessment of trait levels beyond the noise of observed scores, which is due to the inherent unreliability of psychological measures. Overall, the results indicated that the changes detected by full-length measures across waves were almost always captured by CATs. The agreement between full-length measures and CATs in detecting score changes was approximately 80% (ranging from 75.4 to 89.87%), with even higher percentages observed for diagnostic categorizations (ranging from 81.16 to 95.78%).

The advantage of ML-MT-based CATs becomes even more compelling when compared to the performance of short static measures. This comparison was made considering the depression scale. Specifically, the performance of the full-length 9-item version of the instrument (PHQ-9) was compared with that of the short static 2-item version (PHQ-2). The results indicated that the ML-MT CAT was significantly more consistent in detecting changes in agreement with the full-length scale than the short static version. This larger agreement was evident in both score fluctuations (PHQ-2 = 61.96%; ML-based CAT = 75.4%; χ^2 (1) = 12.3, $p \leq .001$) and diagnostic categorizations (PHQ-2 = 81.16%; ML-based CAT = 87.8%; $\chi^2(1) = 5.59$, $p < .05$).

This study makes several important contributions to the field of psychological assessment. First, the work highlights the utility of ML-MT-based CATs in psychological testing for various purposes, including diagnosis, screening, and assessment of disorder severity and trait levels of mental health variables. Second, it demonstrates the good performance of ML-MT-based CATs in detecting changes in psychological variables collected at different time points, which exceeds that of established short static scales. From a clinical perspective, by shortening test administration times and reducing the burden associated with the testing process for both respondents and professionals, ML-MT-based CATs could facilitate the assessment of multiple variables across multiple testing sessions[2]. This allows for more comprehensive screening and frequent monitoring of mental health. According to the literature, since CATs are faster to administer and do not present respondents with the same set of questions each time, they could also reduce boredom, make the testing process more engaging, and consequently lead to more accurate and high-quality responses[3–5]. In addition, ML-MT-based CATs could foster greater interest and motivation in the testing process while reducing the risk of biased responses due to memory effects. However, these aspects were not directly assessed in the current study. Nevertheless, a tangible result that emerged from the study is the ability of CATs to provide accurate assessments of trait levels, capable of detecting changes with an accuracy comparable to that of the full-length test versions, but with considerably greater efficiency. From a research perspective, conducting more rapid evaluations offers a greater opportunity to monitor the effectiveness of treatment programs in real-world settings. This allows researchers to collect more data, thereby improving the rigor of research programs[1].

A notable aspect of interest regarding the results of this study is their methodological robustness. The results were obtained through a cross-validation approach, implying that the performance of the developed CAT procedures was verified on data different from those used to train the algorithms. However, a limitation of this study pertains to the sample size used for training the algorithms. In fact, while the sample size was sufficient, employing larger datasets could potentially have led to even better performance. Future research should investigate this possibility by replicating our findings with larger samples. Furthermore, future research should also try to integrate external variables such as age, gender, and comorbidities into the assessment procedures. This integration may further improve the efficiency and accuracy of the algorithms. Indeed, previous research in the field of CAD has demonstrated that these variables play a pivotal role in enhancing the level of personalization, efficiency, and accuracy achievable through ML-based CATs[41].

The current study did not identify any specific patterns in the performance of the method related to gender, age, number of items presented, or psychological variables. However, it is possible that other individual differences may affect the performance of ML-based CATs in terms of accuracy or efficiency for specific groups. Further research could investigate these factors to enhance the capabilities of the proposed method. It is important to note that the findings of this study, although robust and promising, are based on a non-clinical sample predominantly composed of young individuals. It is recommended that future studies extend and potentially refine these results by including clinical and more diverse samples. Furthermore, future research should investigate the efficacy of the proposed method in long-term studies conducted in real-world settings.

As a further note, it is important to mention that the time frame considered for the PHQ-9 was modified from the original version of the instrument, which set it at two weeks, to a period of six months. This change was made to align the time frame of the depression measure with that of the other collected anxiety measures, which is six months according to the DSM-V. Moreover, the temporal framework was modified to align with the interval between the four waves of data collection. Although this adjustment does not affect the usefulness of the observed results or the feasibility of the proposed method, future studies are needed to confirm and extend the results of the present work to the original version of the PHQ-9. Furthermore, future research should extend the results by considering the performance of the proposed method when it is intended to predict a diagnosis based on criteria external to the test, such as professional diagnoses or results of other diagnostic tests. In this regard, some studies have demonstrated the usefulness of ML-based methods in terms of diagnostic accuracy and administration efficiency even when applied to the traditional PHQ-9 and in relation to external diagnostic criteria[22,41]. However, further research is needed to investigate the ability of ML-based methods to detect changes in these directions.

In a concluding remark, it is important to cite some ethical implications associated with the application of ML in the domain of mental health. ML algorithms have the potential to perpetuate existing biases in their training data, which may consequently result in unfair treatment of certain demographic groups[76]. Ensuring transparency in the decision-making processes of these algorithms is essential to improve algorithm outcomes and maintain trust between patients and mental health professionals. To adequately address these concerns, there is a need to engage in thoughtful deliberation and establish sound ethical guidelines governing the application of ML-based methods in mental health settings.

Conclusion

In recent years, the exploration of innovative methods in various scientific disciplines, including healthcare, marketing, finance and engineering, has demonstrated the potential to address important contemporary challenges[77–82]. This paper focused specifically on the novel application of ML-based methods in the field of psychodiagnostic assessment. In particular, the work demonstrated that ML-MT-based CATs can provide trait-level estimates comparable to those of full-length tests, while requiring more than 50% fewer items to achieve this goal (from 54.1 to 56.5%).

In addition, this work demonstrated the ability of these instruments to detect changes in trait levels and diagnostic categorization over time, closely matching the results of full-length tests. This capability to track and detect change is highly valuable in clinical settings, highlighting how these tools can be effectively used for patient monitoring, treatment evaluation, and research programs.

This study illustrated the potential of ML-MT-based CATs for efficient and accurate mental health assessment, providing an alternative to traditional methods. Indeed, although IRT-based CATs are valued for their personalized and efficient assessments, they require a significant initial investment in item calibration and validation[40], and their underlying assumptions are not always applicable to all psychodiagnostic contexts[43,83]. The ML-based approach overcomes these limitations and offers a promising direction for the dynamic monitoring of psychological change.

Data availability

The database analyzed in this study was described in Vidal-Arenas et al. (2023) and is publically available at https://osf.io/jz4ge/?view_only=ea57acad6e6a4a6e86fa5a57caf493c.

Received: 30 July 2024; Accepted: 18 February 2025
Published online: 22 February 2025

References

1. Gibbons, R. D., Weiss, D. J., Frank, E. & Kupfer, D. Computerized adaptive diagnosis and testing of mental health disorders. *Annu. Rev. Clin. Psychol.* **12**, 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634 (2016).
2. Gibbons, R. D. et al. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr Serv.* **59**, 361–368. https://doi.org/10.1176/ps.2008.59.4.361 (2008).
3. Gass, C. S. Use of the MMPI-2 in neuropsychological evaluations. In (ed Butcher, J. N.) *Oxford Handbook of Personality Assessment* (pp. 432–456). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195366877.013.0023
4. McHorney, C. A. & Tarlov, A. R. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual. Life Res.* **4**, 293–307. https://doi.org/10.1007/BF01593882 (1995).
5. Rose, M., Bjorner, J. B., Becker, J., Fries, J. F. & Ware, J. E. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported outcomes measurement information system (PROMIS). *J. Clin. Epidemiol.* **61**, 17–33. https://doi.org/10.1016/j.jclinepi.2006.06.025 (2008).
6. Means-Christensen, A. J., Sherbourne, C. D., Roy-Byrne, P. P., Craske, M. G. & Stein, M. B. Using five questions to screen for five common mental disorders in primary care: diagnostic accuracy of the anxiety and depression detector. *Gen. Hosp. Psychiatry.* **28**, 108–118. https://doi.org/10.1016/j.genhosppsych.2005.08.010 (2006).
7. Kroenke, K., Spitzer, R. L. & Williams, J. B. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x (2001).
8. Kroenke, K., Spitzer, R. L., Williams, J. B. & Löwe, B. An ultra-brief screening scale for anxiety and depression: the PHQ–4. *Psychosomatics* **50**, 613–621. https://doi.org/10.1016/S0033-3182(09)70864-3 (2009).
9. Khubchandani, J., Brey, R., Kotecki, J., Kleinfelder, J. & Anderson, J. The psychometric properties of PHQ-4 depression and anxiety screening scale among college students. *Arch. Psychiatr Nurs.* **30**, 457–462. https://doi.org/10.1016/j.apnu.2016.01.014 (2016).
10. Löwe, B. et al. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. *Med. Care.* **46**, 266–274. https://doi.org/10.1097/MLR.0b013e318160d093 (2008).
11. Cuijpers, P., Smits, N., Donker, T., Ten Have, M. & de Graaf, R. Screening for mood and anxiety disorders with the five-item, the three-item, and the two-item mental health inventory. *Psychiatry Res.* **168**, 250–255. https://doi.org/10.1016/j.psychres.2008.05.012 (2009).

12. Goldberg, D. P. et al. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychol. Med.* **27**, 191–197. https://doi.org/10.1017/S0033291796004242 (1997).
13. Gaynes, B. N. et al. Feasibility and diagnostic validity of the M-3 checklist: a brief, self-rated screen for depressive, bipolar, anxiety, and post-traumatic stress disorders in primary care. *Ann. Fam Med.* **8**, 160–169. https://doi.org/10.1370/afm.1092 (2010).
14. Connor, K. M., Davidson, J. R. & SPRINT A brief global assessment of post-traumatic stress disorder. *Int. Clin. Psychopharmacol.* **16**, 279–284. https://doi.org/10.1097/00004850-200109000-00005 (2001).
15. Lang, A. J. & Stein, M. B. An abbreviated PTSD checklist for use as a screening instrument in primary care. *Behav. Res. Ther.* **43**, 585–594. https://doi.org/10.1016/j.brat.2004.04.005 (2005).
16. Herr, D. J. & Buchanan, E. M. Generativity and other buffers of death awareness in first responders. *Anxiety Stress Coping.* **33**, 193–206. https://doi.org/10.1080/10615806.2019.1695522 (2020).
17. Tiet, Q. Q., Schutte, K. K. & Leyva, Y. E. Diagnostic accuracy of brief PTSD screening instruments in military veterans. *J. Subst. Abus Treat.* **45**, 134–142. https://doi.org/10.1016/j.jsat.2013.01.010 (2013).
18. Shields, R. E. et al. Brief mental health disorder screening questionnaires and use with public safety personnel: a review. *IJRPH* **18**, 3743. https://doi.org/10.3390/ijerph18073743 (2021).
19. Prieto, L., Alonso, J. & Lamarca, R. Classical test theory versus Rasch analysis for quality-of-life questionnaire reduction. *Health Qual. Life Outcomes.* **1**, 27. https://doi.org/10.1186/1477-7525-1-27 (2003).
20. Muñiz, J. Test theories: classical theory and item response theory. *Pap Psicol.* **31**, 57–66 (2010).
21. Xu, L., Jiang, Z., Han, Y., Liang, H. & Ouyang, J. Developing computerized adaptive testing for a National health professionals' exam: an attempt from psychometric simulations. *Perspect. Med. Educ.* **12**, 462–471. https://doi.org/10.5334/pme.855 (2023).
22. Colledani, D., Anselmi, P. & Robusto, E. Machine learning-decision tree classifiers in psychiatric assessment: an application to the diagnosis of major depressive disorder. *Psychiatry Res.* **322**, 115127. https://doi.org/10.1016/j.psychres.2023.115127 (2023).
23. Gonzalez, O. Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychol. Methods.* **26**, 236–254. https://doi.org/10.1037/met0000317 (2021).
24. Gonzalez, O. Psychometric and machine learning approaches to reduce the length of scales. *Multivar. Behav. Res.* **56**, 903–919. https://doi.org/10.1080/00273171.2020.1781585 (2021).
25. Gibbons, R. D. & deGruy, F. V. Without wasting a word: extreme improvements in efficiency and accuracy using computerized adaptive testing for mental health disorders (CAT-MH). *Curr. Psychiatry Rep.* **21**, 1–9. https://doi.org/10.1007/s11920-019-1053-9 (2019).
26. Graham, A. K. et al. Validation of the computerized adaptive test for mental health in primary care. *Ann. Fam Med.* **17**, 23–30. https://doi.org/10.1370/afm.2316 (2019).
27. Chang, H. H. Psychometrics behind computerized adaptive testing. *Psychometrika* **80**, 1–20. https://doi.org/10.1007/S11336-014-9401-5 (2015).
28. Thompson, N. A. Item selection in computerized classification testing. *Educ. Psychol. Meas.* **69**, 778–793. https://doi.org/10.1177/0013164408324460 (2009).
29. Thompson, N. A. A practitioner's guide for variable-length computerized classification testing. *PARE* **12**, 1–13. https://doi.org/10.7275/fq3r-zz60 (2019).
30. De Ayala, R. J. *The Theory and Practice of Item Response Theory (Second edition)* (The Guilford Press, 2022).
31. Hambleton, R. K., Swaminathan, H. & Rogers, H. J. *Fundamentals of Item Response Theory* (SAGE, 1991).
32. Loken, E. & Rulison, K. L. Estimation of a four-parameter item response theory model. *Br. J. Math. Stat. Psychol.* **63**, 509–525. https://doi.org/10.1348/000711009X474502 (2010).
33. Van der Linden, W. J. & Glas, C. A. Computerized adaptive testing: theory and practice. *Comput. Adapt. Testing: Theory Pract.* https://doi.org/10.1007/0-306-47531-6 (2000).
34. Wainer, H. et al. *Computerized Adaptive Testing: A Primer* (Lawrence Erlbaum Associates, Inc., 1990).
35. Cella, D. et al. The Patient-Reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. *Med. Care.* **45**, S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55 (2007).
36. Kubiszyn, T. & Borich, G. D. *Educational Testing and Measurement* (Wiley, 2024).
37. Pilkonis, P. A. et al. Item banks for measuring emotional distress from the Patient-Reported outcomes measurement information system (PROMIS®): depression, anxiety, and anger. *Assessment* **18**, 263–283. https://doi.org/10.1177/1073191111411667 (2011).
38. Reise, S. P. & Waller, N. G. Item response theory and clinical measurement. *Annu. Rev. Clin. Psychol.* **5**, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553 (2009).
39. Simms, L. J. & Clark, L. A. Validation of a computerized adaptive version of the schedule for nonadaptive and adaptive personality (SNAP). *Psychol. Assess.* **17**, 28–43. https://doi.org/10.1037/1040-3590.17.1.28 (2005).
40. Delgado-Gomez, D., Laria, J. C. & Ruiz-Hernandez, D. Computerized adaptive test and decision trees: A unifying approach. *Expert Syst. Appl.* **117**, 358–366. https://doi.org/10.1016/j.eswa.2018.09.052 (2019).
41. Colledani, D., Robusto, E. & Anselmi, P. Shortening and personalizing psychodiagnostic assessments with decision tree-machine learning classifiers: an application example based on the patient health questionnaire-9. *Int. J. Ment Health Addict.* https://doi.org/10.1007/s11469-024-01332-x (2024).
42. Gibbons, R. D., Chattopadhyay, I., Meltzer, H. Y., Kane, J. M. & Guinart, D. Development of a computerized adaptive diagnostic screening tool for psychosis. *Schizophr Res.* **245**, 116–121. https://doi.org/10.1016/j.schres.2021.03.020 (2022).
43. Yan, D., Lewis, C. & Stocking, M. Adaptive testing with regression trees in the presence of multidimensionality. *J. Educ. Behav. Stat.* **29**, 293–316. https://doi.org/10.3102/10769986029003293 (2004).
44. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification and regression trees. *Classif. Regres. Trees.* https://doi.org/10.1201/9781315139470 (2017).
45. Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. *Data Mining: Practical Machine Learning Tools and Techniques* (2016).
46. Criminisi, A., Shotton, J. & Konukoglu, E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found. Trends Comput. Graph Vis.* **7**, 81–227. https://doi.org/10.1561/0600000035 (2012).
47. Gupta, B., Rawat, A., Jain, A., Arora, A. & Dhami, N. Analysis of various decision tree algorithms for classification in data mining. *Int. J. Comput. Appl.* **163**, 15–19. https://doi.org/10.5120/ijca2017913660 (2017).
48. Gray, R. M. *Entropy and Information Theory* (Springer, 2011).
49. Lu, F. & Petkova, E. A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Stat. Med.* **33**, 401–421. https://doi.org/10.1002/sim.5937 (2014).
50. McArdle, J. J. Adaptive testing of the Number Series Test using standard approaches and a new decision tree analysis approach. In *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences* (2013). https://doi.org/10.4324/9780203403020-22
51. Ueno, M. & Songmuang, P. Computerized adaptive testing based on decision tree. In *Proceedings – 10th IEEE International Conference on Advanced Learning Technologies, ICALT 2010* (2010).
52. Devine, J. et al. Evaluation of computerized adaptive tests (CATs) for longitudinal monitoring of depression, anxiety, and stress reactions. *J. Affect. Disord.* **190**, 846–853. https://doi.org/10.1016/j.jad.2014.10.063 (2016).
53. Kim-Kang, G. & Weiss, D. J. Comparison of computerized adaptive testing and classical methods for measuring individual change. In *Proceedings of the GMAC Conference on Computerized Adaptive Testing* (2007). (2007). Available from www.psych.umn.edu/psylabs/CATCentral

54. Grunebaum, M. F., Mann, J. J., Galfalvy, H. C. & Gibbons, R. D. Computerized-adaptive vs. traditional ratings of depression and suicidal thoughts: an assay sensitivity pilot study in a ketamine clinical trial. *Front. Psychiatry*. **12**, 602976. https://doi.org/10.3389/fpsyt.2021.602976 (2021).
55. Vidal-Arenas, V., Ortet-Walker, J., Ibáñez, M. I., Ortet, G. & Mezquita, L. Self-reported DSM-5 anxiety severity measures: evidence of validity and reliability in Spanish youths. *Psicothema* **33**, 312–319. https://doi.org/10.7334/psicothema2020.398 (2021).
56. Lebeau, R. T. et al. G. A dimensional approach to measuring anxiety for DSM-5. *Int. J. Methods Psychiatr Res.* **21**, 258–272. https://doi.org/10.1002/mpr.1369 (2012).
57. Meyerhoff, J. et al. Evaluation of changes in depression, anxiety, and social anxiety using smartphone sensor features: longitudinal cohort study. *J. Med. Internet Res.* **23**, e22844. https://doi.org/10.2196/22844 (2021).
58. Moreno-Agostino, D. et al. Global trends in the prevalence and incidence of depression: a systematic review and meta-analysis. *J. Affect. Disord*. **281**, 235–243. https://doi.org/10.1016/j.jad.2020.12.035 (2021).
59. Jefferies, P. & Ungar, M. Social anxiety in young people: a prevalence study in seven countries. *PLoS One*. **15**, e0239133. https://doi.org/10.1371/journal.pone.0239133 (2020).
60. Xiong, P., Liu, M., Liu, B. & Hall, B. J. Trends in the incidence and dalys of anxiety disorders at the global, regional, and National levels: estimates from the global burden of disease study 2019. *J. Affect. Disord*. **297**, 83–93. https://doi.org/10.1016/j.jad.2021.10.022 (2022).
61. OECD/European Union. *Health at a Glance: Europe 2018: State of Health in the EU Cycle* (OECD Publishing, 2018). https://doi.org/10.1787/health_glance_eur-2018-en
62. Vidal-Arenas, V. et al. Longitudinal measurement invariance of the DSM-5 anxiety and depression severity measures. *Eur. J. Psychol. Assess.* https://doi.org/10.1027/1015-5759/a000791 (2023).
63. Beesdo-Baum, K., Klotsche, J., Knappe, S., Craske, M. G., LeBeau, R. T., Hoyer, J.,… Wittchen, H. U. Psychometric properties of the dimensional anxiety scales for DSM-V in an unselected sample of German treatment-seeking patients. Depress. Anxiety29, 1014–1024 (2012). https://doi.org/10.1002/da.21994.
64. Costantini, L. et al. Screening for depression in primary care with patient health Questionnaire-9 (PHQ-9): A systematic review. *J. Affect. Disord*. **279** https://doi.org/10.1016/j.jad.2020.09.131 (2021).
65. Manea, L., Gilbody, S. & McMillan, D. Optimal cut-off score for diagnosing depression with the patient health questionnaire (PHQ-9): A meta-analysis. *CMAJ* 184. https://doi.org/10.1503/cmaj.110829 (2012).
66. Spitzer, R. L., Kroenke, K., Williams, J. B. & Patient Health Questionnaire Primary Care Study Group. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA* **282**, 1737–1744. https://doi.org/10.1001/jama.282.18.1737 (1999).
67. Kroenke, K., Spitzer, R. L. & Williams, J. B. The patient health Questionnaire-2: validity of a two-item depression screener. *Med. Care*. **41**, 1284–1292. https://doi.org/10.1097/01.MLR.0000093487.78664.3C (2003).
68. Hall, M. et al. The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.* **11**, 10–18. https://doi.org/10.1145/1656274.1656278 (2009).
69. Wang, Y. & Witten, I. H. Induction of model trees for predicting continuous classes. In *Proceedings of the 9th European Conference on Machine Learning* (1997).
70. Quinlan, J. R. Learning with continuous classes. In *Australian Joint Conference on Artificial Intelligence* (1992).
71. Patange, A. D. & Jegadeeshwaran, R. A machine learning approach for vibration-based multipoint tool insert health prediction on vertical machining centre (VMC). *Measurement* **173**, 108649. https://doi.org/10.1016/j.measurement.2020.108649 (2021).
72. Mansour, Y. Pessimistic decision tree pruning based on tree size. In *Proc. 14th International Conference on Machine Learning* (1997).
73. Steiger, J. H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **87**, 245–251. https://doi.org/10.1037/0033-2909.87.2.24 (1980).
74. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Erlbaum, 1988).
75. Lenhard, W. & Lenhard, A. *Computation of effect sizes.* Retrieved from: https://www.psychometrica.de/effect_size.html. Psychometrica. (2022). https://doi.org/10.13140/RG.2.2.17823.92329
76. Agarwal, R. et al. Addressing algorithmic bias and the perpetuation of health inequities: an AI bias aware framework. *Health Policy Techn*. **12**, 100702. https://doi.org/10.1016/j.hlpt.2022.100702 (2023).
77. Dixon, M. F., Halperin, I. & Bilokon, P. *Machine Learning in Finance*Vol. 1170 (Springer, 2020). https://doi.org/10.1007/978-3-030-41068-1International Publishing.
78. Dzyabura, D. & Yoganarasimhan, H. Machine learning and marketing. In N. Mizik & D. M. Hanssens (Eds.), *Handbook of Marketing Analytics* (pp. 255–279). Edward Elgar Publishing (2018). https://doi.org/10.4337/9781784716752.00023
79. Eid, N., Yosri, N., El-Seedi, H. R., Awad, H. M. & Emam, H. E. Ag@ Sidr honey nanocomposite: chemical profiles, antioxidant and microbicide procurator. *Biocatal. Agric. Biotechnol.* **51**, 102788. https://doi.org/10.1016/j.bcab.2023.102788 (2023).
80. Philip, A. K., Samuel, B. A., Bhatia, S., Khalifa, S. A. & El-Seedi, H. R. Artificial intelligence and precision medicine: a new frontier for the treatment of brain tumors. *Life* **13**, 24. https://doi.org/10.3390/life13010024 (2022).
81. Reich, Y. Machine learning techniques for civil engineering problems. *Comput. Aided Civil Infrastructure Eng.* **12**, 295–310. https://doi.org/10.1111/0885-9507.00065 (1997).
82. Sarker, I. H. Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 160. https://doi.org/10.1007/s42979-021-00592-x (2021).
83. Anselmi, P. et al. An item response theory-based scoring of the South Oaks gambling screen–revised adolescents. *Assessment* **29**, 1381–1391. https://doi.org/10.1177/10731911211017657 (2022).

Acknowledgements

Thanks to Dr. T. Santippe for her invaluable presence and support.

Author contributions

D.C. conceptualized the work, developed the methodology and carried out the analyses. D.C. and P.A. interpreted the results of the analyses, and wrote the main manuscript text. D.C., P.A., and C.B. reviewed the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to P.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025