Received: 7 November 2023 | Revised: 6 May 2024 | Accepted: 29 June 2024

DOI: 10.1002/mpr.2032

WILEY

ORIGINAL ARTICLE

Developing an accurate and efficient tool for the internalizing spectrum: A simulation study of the adaptive algorithm to the Inventory of Depression and Anxiety Symptoms II (IDAS-II)

M. Sanchez-Garcia[1,2] | A. De la Rosa-Cáceres[1,2] | G. Rossi[3] | C. Diaz-Batanero[1,2]

[1]Department of Clinical and Experimental Psychology, University of Huelva, Huelva, Spain

[2]University of Huelva, Research Center for Natural Resources, Health and the Environment, Huelva, Spain

[3]Department of Psychology, Personality and Psychopathology Research Group, Vrije Universiteit Brussel, Brussels, Belgium

Correspondence

C. Diaz-Batanero, Department of Clinical and Experimental Psychology, Facultad de Ciencias de la Educación, Research Center for Natural Resources, Health and the Environment, University of Huelva, Huelva 21071, Spain.
Email: carmen.diaz@dpsi.uhu.es

Funding information

Ministerio de Ciencia e Innovación (MICIU/AEI/10.13039/501100011033), Grant/Award Number: PID2020-116187RB-I00; Ministerio de Universidades, Grant/Award Number: FPU19/00144

Abstract

Objectives: This research simulates an adaptive version of the IDAS-II (IDAS-CAT).

Methods: 2021 participants from both community ($n = 1692$) and patients ($n = 329$) samples completed the IDAS-II. Item response theory metric properties of the IDAS-II full test and the 20-items of the general depression (GD) scale were obtained. The efficiency and accuracy of different computerized adaptive algorithms were simulated. Different subsamples completed additional external measures in order to gather evidence of validity of the scores estimated with the simulated adaptive algorithms selected.

Results: Both unidimensional computerized adaptive testing algorithm selected for the GD scale and the bifactor model chosen for the full test, allow 70% reduction in the length of administration, maintaining a measurement error below 0.30 on the general and 0.50 on the specific factors. Results show high correlations of the scores estimated with the adaptive algorithms and the estimates based on the full test, as well as correlations with external criteria almost equal to those generated with the full test.

Conclusions: IDAS-CAT could be a reliable and fast tool for measuring internalizing spectrum.

KEYWORDS

adaptive testing, anxiety, assessment, depression, IDAS-II, internalizing

1 | BACKGROUND

Scientific evidence consistently reflects high comorbidity among internalizing problems (Kessler et al., 2015; Lamers et al., 2011). These observations have led to the development of alternative explanatory models and reformulations of the diagnostic categorical taxonomies towards dimensional and transdiagnostic approaches. Initially, the results of this extensive line of research on the structure of these disorders led to the proposal of a general latent liability, called *internalizing spectrum*, which would explain the co-occurrence of these disorders (Krueger & Markon, 2006). The definition of the internalizing spectrum has guided scientific advances in this vein for

decades (Achenbach et al., 2016). More recently, the Hierarchical Taxonomy of Psychopathology (HiTOP) (Kotov et al., 2017), has broaden this conceptualization to general psychopathology, grouping symptoms into higher dimensions, that are organized hierarchically (symptoms group into disorders, spectra and ultimately into higher order factors). In addition to this taxonomic proposal, aimed at reorienting the diagnosis of disorders, transdiagnostic intervention proposals have also been developed in recent years. An example of this is the Unified protocol, which has developed specific proposals for the intervention of emotional disorders (Barlow et al., 2011, 2016, 2017).

In order to both promote further research in the field of internalizing disorders, as well as to implement these recent taxonomic and intervention proposals in the clinical context, and screen for individual in need for treatment in community samples, it is necessary to develop appropriate measurement tools. For the measurement of internalizing problems, the IDAS-II (De la Rosa-Cáceres et al., 2020; Watson et al., 2012) has proven to be a useful tool both in the research field and in applied and clinical settings. Specifically, the IDAS-II consists of 18 nonoverlapping scales assessing internalizing symptoms, allowing a more congruent evaluation with recent transdiagnostic approaches. It also includes a general depression (GD) scale, containing items from other IDAS-II scales that gives an overall depression symptom score. The operational definition of this instrument and its subscales is theoretically framed within the internalizing spectrum (Watson et al., 2012), which includes the sub-factors of fear, distress, and mood.

While the traditional application of this test has shown great interest and utility, its length may limit its application in research or clinical and applied settings with limited administration time or can lead to respondent burden when administering larger batteries (Yates & Taub, 2003). In addition, the static and traditional application of a test implies that the same items are applied, in the same order, to all people regardless of their level in the measured variable, so that some items will provide little or no discriminative ability for some people. An increasingly popular alternative is the use of computerized adaptive testing (CAT) based on item response theory (IRT) models (Embretson & Reise, 2000; Feuerstahler et al., 2020).

In a CAT procedure, instead of applying all (or the same) items to all persons, using computerized IRT algorithms, each person will be given the items that provide the most information for his or her level in the variable previously estimated. For the development of a CAT it is necessary to have a larger bank of items previously calibrated (Wainer, 2000). When applying a CAT, after the response to the first item, the person's level in the construct is estimated (van der Linden & Pashley, 2000). From this estimate, the item that provides the most information (precision) of the rest of items is chosen and, after the response, the score is re-estimated. This process continues until a certain stopping criterion is reached. To test the efficiency and usefulness of CAT procedures, it is common practice to perform computer simulations of item banks based on a series of assumptions (e.g., Stochl et al., 2016; Tsaousis et al., 2021). Thus, these simulations study the accuracy and efficiency of the adaptive algorithm with different starting conditions, estimation of people's scores and stopping criteria, and provide evidence of validity of the scores derived with the CAT algorithm (Barnard, 2018; Han & Kosinki, 2014).

The simulation, development and evaluation of CAT procedures has become widespread in recent years in the field of psychopathology and drug use (Guinart et al., 2021; Hulvershorn et al., 2022; Mustanski et al., 2021; Sunderland et al., 2017). These studies showed a saving in the number of items needed in the assessment, while maintaining adequate accuracy and agreement with the full measure. To date, most of these adaptive applications have been performed on unidimensional models (e.g., Sunderland et al., 2017). IRT models were applied to capture the full continuum of severity along the dimension with a more efficient smaller subset of items (Embretson & Reise, 2000). However, when high comorbidity among symptoms is observed, as in internalizing disorders, it is more realistic and useful to use multidimensional approaches (Gibbons et al., 2016; Reckase, 2009). Specifically, bifactor IRT models allow estimating the metric properties of items with respect to a general factor and different specific factors as well (Reise et al., 2007; Toland et al., 2017). CAT simulation studies with bifactor models conducted to date with instruments linked to internalizing symptoms show adequate performance and increased efficiency. However, these studies either use measures restricted to only one disorder (Gibbons et al., 2012, 2014) or larger item banks from different instruments designed with different metrics and time-frame (Sunderland et al., 2019). The simulation study of a CAT for the IDAS-II would optimize the assessment of a wide range of symptoms of internalizing disorders, evaluated with the same response format and within the most recent theoretical approach of HiTOP. In addition, the studies conducted to date use patient-specific samples (Gibbons et al., 2008; Gibbons et al., 2012, 2014) or community samples (Sunderland et al., 2019). In this sense, the inclusion of mixed samples would allow a more adequate representation of the entire continuum to be evaluated.

Therefore, the main objective of this work is to show the metric properties of an adaptive version of the IDAS-II test (IDAS-CAT) through a simulation process departing from data collected on a large mixed sample. To this end, we aim to: (1) analyze the uni- and multidimensional structure of the supra-facet level of IDAS-II found in previous works; (2) analyze the metric properties of the items from a unidimensional IRT model with the GD scale and a bifactor model with the 99 items of IDAS-II, including: a general factor (internalizing) and three specific factors (distress, fear and mood); (3) simulate the efficiency and accuracy of a computerized and adaptive version of the IDAS-II (IDAS-CAT) to locate the evaluated individuals scores along a continuum of underlying internalizing problem levels; (4) provide evidence of validity of the scores estimated in this simulation process from the correlations of these scores with the scores of the original IDAS-II scales and the scores on external variables collected in the sample. According to previous literature, it is expected: (1) that, in both confirmatory factor analysis (CFA) models, the 20 GD items fit a unidimensional model and that the 99 IDAS-II items fit a

bifactor model with one general factor (internalizing) and three specific factors (distress, fear, and mood); (2) that the 20 GD items fit a unidimensional IRT model and the 99 IDAS-II items fit a multidimensional, bifactor IRT model, as previously defined. Regarding the simulation process, we expect to (3) that, in both cases, it will be possible to find an adaptive algorithm that allows a significant reduction in the number of items needed to make an adequate measurement; (4) that the scores generated through the selected adaptive algorithms will be accurate (in terms of low estimation error) and efficient (in terms of a significant reduction in the number of items used); (5) that the scores generated through the selected adaptive algorithms will provide evidence of validity based on the relationship with other external variables very close to those provided by the scores obtained in a non-adaptive application of all items. To this end we will examine correlations with scales (i.e. measuring depression, anxiety, post-traumatic stress, obsessive-compulsive symptoms, pathological traits and functional impairment) that have well-established relations with IDAS-II scales (see e.g. De la Rosa et al., 2023; Vittengl et al., 2023).

2 | MATERIAL AND METHODS

2.1 | Participants and procedure

To decide on the proper sample size, representativeness and variability criteria were taken into account. Considering the criterion of representativeness of the Spanish population, a sampling error of 3%, and a confidence interval of 95%, the required sample size would be 1048 participants. According to Ferrando et al. (2022) and Jiang et al. (2016) this sample size would enough to carry out CFA with the intended indicators and factors and to ensure stability on the estimation of IRT parameters. To ensure variability in responses, there were added an additional sample of patients that allow a more adequate representation of the entire continuum to be evaluated.

The final mixed sample was composed of patients and people from the community population (n = 2021) was gathered by three waves of data collection: Wave (1) 620 community adults selected by means of non-probability sampling in the province of Huelva (Spain); Wave (2) 1072 community adults recruited by stratified random sampling, proportionally represented in the Spanish population according to age group, sex, and geographic area. Wave (3) 329 patients from public and private mental health services in the province of Huelva (Spain). Inclusion criteria for community samples were being between 18 and 80 years old and not having a diagnose of any mental disorder. Inclusion criteria for clinical sample were being between 18 and 80 years old and being under treatment in a mental health service during the data collection. Those who met any of the following characteristics were excluded from all samples: having been diagnosed with a medical or psychological disorder that disqualified them from taking the tests, or not signing the informed consent form.

Community adults from wave 2 completed the instruments in an online format, whilst a psychologist administered the questionnaires in paper and pencil format to the wave 1 and 3. The latter completed the measures in rooms set up in the centers where they were recruited. Results from online and paper-and-pencil collection result into quantitative equivalence and this can be both used (Weigold et al., 2013). All participants were informed about the anonymous and voluntary nature of their participation in the study and gave their written informed consent. This study has the approval of the Bioethics Committee of Biomedical Research of Andalusia (Spain) (file number PI 040/18).

2.2 | Measures

All participants completed the *IDAS-II* (Watson et al., 2012; Spanish version by De la Rosa-Cáceres et al., 2020). The IDAS-II consists of 99 items with a Likert-type response format (1 = *not at all* to 5 = *extremely*) (De la Rosa-Cáceres et al., 2020; Kotov et al., 2017) organized on 18 nonoverlapping scales (Dysphoria, Lassitude, Insomnia, Suicidality, Appetite Loss, Appetite Gain, Well-Being, Ill-Temper, Mania, Euphoria, Panic, Social Anxiety, Claustrophobia, Traumatic Intrusions, Traumatic Avoidance, Checking, Ordering, Cleaning) and an overlapping scale made up of 20 items from other scales (GD). Internal consistency in this study ranged from Cronbach's alpha value of 0.75 for Ordering and 0.91 for Dysphoria and Panic. The GD scale reached an alpha value of 0.93.

In order to gather evidence of validity of the scores estimated with the adaptive algorithms selected, the following external measures were applied to different subsamples of the complete group of participants. The number of participants who completed each one can be read in Tables 3 and 5.

Beck Depression Inventory-II (BDI-II; Beck et al., 1996; Spanish version by Sanz et al., 2003). This instrument evaluates the presence and severity of depressive symptomatology in the last two weeks. In this work, we have found high internal consistency values (Cronbach's alpha = 0.90).

Beck Anxiety Inventory (BAI; Beck & Steer, 1990; Spanish version by Sanz & Navarro, 2003). A 21-item inventory that assesses the severity of anxiety symptoms in the last seven days. The BAI showed high internal consistency in this study (Cronbach's alpha value of 0.93).

Post-traumatic Stress Disorder Checklist-Civilian Version (PCL-C; Weathers et al., 1993; Spanish version by Orlando & Marshall, 2002) was used to measure post-traumatic stress disorder (PTSD) symptoms corresponding to criteria B (re-experiencing), C (avoidance), and D (arousal) of the diagnostic criteria for DSM-IV PTSD. In this work we will only use the global score (PCL-C Total; Cronbach's alpha = 0.93).

Obsessive-Compulsive Inventory-Revised (OCI-R; Foa et al., 2002; Spanish version by Fullana et al., 2005). Formed by 18 items measuring the severity of obsessive-compulsive disorder (OCD) symptoms, in six subscales with three items each (Washing, Checking, Ordering, Neutralizing, Hoarding, and Obsessing) and an overall score (OCI-R Total Cronbach's alpha = 0.90; internal consistency

values of the subscales range from Neutralizing (α = 0.67) to Ordering (α = 0.78)).

Personality Inventory for DSM-5 Short Form (PID-5-SF; Krueger et al., 2012; Maples et al., 2015; Spanish version by Díaz-Batanero et al., 2019). This 100-items instrument assesses 25 personality facets organized into five higher order domains as described in DSM-5 section III in the alternative model for personality disorders. The reliability values of the scores in this study in the five domains measured are: Negative Affect (α = 0.87), Detachment (α = 0.90), Antagonism (α = 0.68), Disinhibition (α = 0.89), Psychoticism (α = 0.76).

WHO Disability Assessment Schedule II (WHODAS 2.0; WHO, 2000; Spanish version by Vázquez-Barquero et al., 2000). This instrument was developed from a set of the International Classification of Functioning, Disability and Health (ICF) items to measure functional impairment. Each item is scored on a 5-point Likert scale (0 = *none* to 4 = *extreme or cannot do*) which grades the difficulty experienced by the participant in performing a given activity. In this sample, the value of the internal consistency of the scores (Cronbach's alpha) is 0.91.

2.3 | Data analysis

Firstly, the data collected (either on paper and pencil or online) were used to test the structure of IDAS-II using CFA and to analyze the metric properties of items using IRT (adjusting the items, estimating item parameters, and calculating true theta scores for individuals). Following that, the simulation of different CAT algorithms was undertaken. Based on the previously estimated IRT parameters of individuals and items, simulated responses from all individuals under the specified conditions were estimated: adaptive simulations with stopping criteria (CAT algorithms 1 to 5), adaptive simulations without stopping criteria (full test), and simulations with a fixed number of items that are not adaptive but random. Finally, collected and simulated scores were correlated to provide evidence of the validity of the estimated algorithms. Each of the analyses are described below.

Confirmatory factor analyses (CFAs): The following models were tested: Model 1 (GD factor), in which 20 items were specified as loading into a single factor (Unidimensional GD; GD); Model 2 (one general factor: Internalizing), where the 99 IDAS-II items are grouped in one dimension (Unidimensional Internalizing); Model 3 (first and second order factors), in which three first-order factors were specified – distress, fear/obsession, and mood – that grouped into a second-order factor (Internalizing); Model 4 (hierarchical three-subfactor, bifactor model), in which a general factor with all items (general factor internalizing), and three specific factors - distress, fear/obsession, and mood -into which certain specific items loaded were specified.

The CFA modeling analyses were conducted using diagonal weighted least squares (DWLS) estimation. The robust DWLS method is recommended in situations where the data are measured in ordinal scale and multivariate normality cannot be assumed (Mindrila, 2010; Rhemtulla et al., 2012). To evaluate goodness of fit for each model, the following quantitative fit statistics were used: the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). Following Hu and Bentler (1999), values of CFI and TLI above 0.95, and values of RMSEA and SRMR below 0.06 and 0.08, respectively, were considered indicative of good fit. In order to compare the models, other fit statistics such as the Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC) will be obtained: smaller values indicate a better model fit.

Item Response Theory Analysis: Two separate IRT analyses will be conducted: one for the unidimensional GD model (20 items) and one for a bifactor model with the same structure as the previously defined model 4. The item parameters and the fit of the IRT models were estimated using a full information item factor analysis procedure using an expectation-maximization (EM) approach (Bock & Aitkin, 1981). Given that our data are essentially ordinal, we used a graded response model (Gibbons et al., 2007; Samejima, 1997), to calculate item parameters and person scores (true theta scores). In the unidimensional model, a one-factor graded response model was used. In the bifactor model we used a multidimensional graded response model. Item fit was assessed using the $S\text{-}X^2$ statistic (Kang & Chen, 2008).

Computerized Adaptive Test Simulations: The IRT parameters estimated previously were used to generate responses to the items in the different dimensions used in the two models analyzed in the previous phase (unidimensional and bifactorial). For both models, different simulations were carried out which differed from each other based on the termination criteria used. The rest of the phases will be the same for all simulations performed.

Start of the simulation and selection of items: Since our goal is to accurately position individuals along a continuum, in the unidimensional model of GD, items will be selected using Fisher's maximum information criterion for the estimated trait levels (Meijer & Nering, 1999). In the bifactor model, items will be adaptively selected using the D-optimally approach, that maximizes the determinant of the information matrix for the estimated trait levels (Seo & Weiss, 2015). However, using the information maximization as the only criterion for selecting items can result on overexposition of some items and underutilization of others. Simulations based solely on maximizing the information function of the scale show higher percentages of reduction in the number of items used, although the evidence of content validity is compromised (Luecht et al., 1998; Yi & Chang, 2003). In present research, we therefore control both the exposure rate of the items and their content, ensuring the proportionality of the dimensions included in the IDAS-II and providing results closer to a real situation. Specially in the case of the multidimensional model, regardless of the amount of information they provide, items from the three specific factors are used, ensuring the representativeness of the contents and the correct interpretation of the scores obtained.

To avoid overexposure of items with highest discriminative ability the 5-4-3-2-1 method of McBride and Martin (1983) was used for the

first five items. That is, selecting one item at random from among the 5 most informative items as the first item, another at random from among the 4 most informative items as the second, and so on up to the 5th item, from which the most informative item is always selected for the last estimated θ level. In addition, in the multidimensional model a content-balancing procedure to ensure the representativeness of all three specific factors in the CAT was used (Chalmers, 2016). Each specific factor was assigned a weight, depending on the number of items that comprise it. In each simulation, a proportional number of items was chosen from each factor (approximately 0.60 for 'dystress', 0.26 for 'fear/obsessions' and 0.13 for 'mood').

Estimation of persons' severity level: The starting point was set as theta = 0. Successive theta values were estimated using Bayesian maximum a posteriori with standard normal priors (Embretson & Reise, 2000).

Stopping criteria: The following stopping criteria were used: (a) minimum number of items to be applied (4 for unidimensional GD and 12 for the bifactor model); and (b) standard error of measurement equal to or less than 0.30 in the unidimensional model and 0.30 (for the general factor) and 0.50 for each of the specific factors in the multidimensional model (Algorithm 1: $SE(\theta) \leq 0.30$ [unidimensional]/ General: $SE(\theta) \leq 0.30$, Specific: $SE(\theta) \leq 0.50$ [bifactorial]).

For extreme values of theta, analogous to previous work (see, for example, Sunderland et al., 2017, 2019), the change in estimated values of theta from one item to another was used as an additional stopping criterion. Thus, four more algorithms were generated: Algorithm 2: Algorithm 1 + $\Delta\theta < 0.005$, that is, the adaptive algorithm terminates when one of the two conditions is met, either the standard error of measurement is less than or equal to 0.3 or the difference in successive estimates of theta is less than 0.005; Algorithm 3: Algorithm 1 + $\Delta\theta < 0.01$; Algorithm 4: Algorithm 1 + $\Delta\theta < 0.05$ years Algorithm 5: Algorithm 1 + $\Delta\theta < 0.10$. In order to compare the results obtained, two more algorithms were run: the first one did not use any termination criteria (full test) and the second one used the same number of items used in the most efficient of the five specified algorithms, but the selection criterion of the items was not adaptive but random.

As measures of the accuracy and efficiency of the adaptive algorithms we calculated: (a) correlations between the thetas estimated in the CAT simulations with the theta scores obtained with the full item bank (full test) and the original theta scores (true theta); (b) the average of the standard errors of the estimated thetas as a measure of the accuracy of each of the simulations; and (c) the average number of items used in each simulation as a measure of the efficiency of the simulation.

Evidence of validity: Correlations between the thetas estimated in the CAT simulations and the full test (i.e. sum scores of original IDAS-II 18 subscales and GD scale) with different external variables: BDI, BAI, OCI, PID-5-SF, PCL-C and WHODAS 2.0.

CFA was conducted with the "cfa" module of the "lavaan" (Rosseel, 2012). IRT analyses were conducted using the *R* package "mirt" (Chalmers, 2012). CAT simulations were performed using the *R* package 'mirtCAT' (Chalmers, 2016). Correlations between simulated scores and full scores and external measures were computed with SPSS 27.0.

2.3.1 | Transparency and openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. This study's design and its analysis were not pre-registered.

3 | RESULTS

3.1 | Confirmatory factor analysis

Table 1 shows the fit for all models tested, the models with best absolute and relative fit indices being the unidimensional GD model (model 1) and the bifactor model (model 4). Thus, these two models will be our input for the CAT simulation. In model 1 the standardized factor loadings range from 0.25 (recoded item 8 of the Well-Being scale) to 0.81 (item 3 of dysphoria). In model 4 the factor loadings

TABLE 1 Fit statistics for confirmatory factor models.

Model	χ^2	df	CFI	TLI	SRMR	RMSEA	90% RMSEA CI	BIC	AIC
Model 1. One factor: General depression (20 items)	**895.13**	**170**	**0.984**	**0.982**	**0.064**	**0.046**	**0.043/.049**	**103,893.65**	**103,669.72**
Model 2. One factor: Internalizing (99 items)	46,213.37	4752	0.917	0.915	0.085	0.067	0.067/.068	506,996.14	505,894.63
Model 3. One higher order factor [internalizing] and three-subfactor [distress, fear, and mood] (99 items)	43,446.21	4749	0.922	0.920	0.082	0.062	0.065/.065	498,505.73	497,387.53
Model 4. Bifactor: Three-subfactor and one general factor of internalizing (99 items)	**22,918.87**	**4653**	**0.963**	**0.962**	**0.061**	**0.045**	**0.045/.046**	**490,718.08**	**489,065.81**

Note: Estimation Method: BIC and AIC were estimated using a Maximum Likelihood (ML) estimator. All chi-square values are statistically significant. Best fitting models are displayed in bold.

Abbreviations: AIC, akaike information criterion; BIC, bayesian information criterion; CFI, comparative fit Index; DWLS, diagonally weighted least squares; RMSEA, root mean square error approximation; SRMR, standardized root mean square residual; TLI, tucker lewis index.

on the general factor (internalizing) range from 0.06 (item 3 of Well-Being) to 0.75 (item five of Mania).

3.2 | Item response theory

Unidimensional GD (20 items): IRT parameters (discrimination [a], thresholds category parameters [b_1, b_2, b_3, b_4], item difficulty (or severity; b_{MEAN} = mean category threshold values), and the fit values for each of the items are shown in Supporting Information S1: Table S1.

The items in general showed adequate values for the discrimination parameter, ranging 0.52–2.90 (average 1.85). 18 of the 20 items had positive difficulty parameters, that is, above the average value with only the two Well-Being items having negative position values. Most extreme position values corresponded with item 8 (from Well-Being) ("He sentido que he tenido mucha energía" [I felt I had a lot of energy]) with bMEAN = −2.82 and item nine on Dysphoria ("He hablado más despacio que lo habitual" [I have spoken more slowly than usual]), b_{MEAN} = 2.51(see Supporting Information S4: Figure S1 in supplementary material). IRT analysis showed that 4 out of 20 items (25%) do not fit ($p < 0.01$).

The difficulty (or severity) item parameters (see Supporting Information S1: Table S1) and test information function (Supporting Information S4: Figure S1) indicated that the items analyzed are useful for discriminating people with medium or high levels of the construct. Supporting Information S4: Figure S1 shows that most of the items were located between values of 0.25 and 1.25, with the dysphoria items being the most centered and the most informative.

3.2.1 | Bifactor model (99 items)

Supporting Information S2: Table S2 shows the parameter values on the general factor of the 99 items used for the bifactor model. 21 of the items (20%) had fit values (S-X^2 *p*-value) below 0.05. Of these, only three have S-X^2 *p*-values below 0.01. The values of the discrimination parameter of the general factor items ranged from 0.12 to 2.87, the central 90% of the discrimination values were between 0.50 and 2.33. Item position parameters ranged from −7.44 (item 3 of Well-Being) to +10.65 (item 7 of Cleaning), with values from 0.45 to 3.37 on the central 90% of the items. Most of the items had medium-high positions, showing these positions the highest values in the information function, thus providing greater precision in the measurement for people with medium-high levels of internalizing.

The contents of Well-Being were located on lower levels of internalizing problems, while the contents of cleaning or euphoria were associated with higher levels. Again, the most discriminative and more centered items were those of Dysphoria, Panic, Appetite Gain and Appetite Loss (see Supporting Information S5: Figure S2).

3.3 | Computerized adaptive test simulations

3.3.1 | Unidimensional general depression (20 items)

The results of the different CAT simulations tested for the unidimensional model are shown in Table 2. All the algorithms contrasted implied a reduction in the number of items used in the estimation of the person's severity level. Using both precision and efficiency criteria, the solution offered by Algorithm 4 seemed the most appropriate. With only six items (reduction of 70% of items) it achieved a precision very close to 0.30 and the theta values obtained were closely correlated with those obtained with the full test. Including the same number of items, the randomized algorithm showed lower precision values and lower correlations with true theta and full scores. The analysis of the measurement error obtained with the algorithm selected, showed that people with lower levels (theta below 0) were measured less accurately, higher standard errors of measurement were obtained, and therefore more items were

TABLE 2 Simulation studies comparing the different termination criteria with the application of the full test or a random selection of items.

	N items				Correlations	
	Mean	% Reduction	Mean θ	Mean SE(θ)	True	Full
True	---	---		---	1.00	0.97
Full	20	0	0.013	0.24	0.97	1.00
Algorithm 1	8	60	0.020	0.30	0.95	0.98
Algorithm 2	8	60	0.020	0.30	0.95	0.98
Algorithm 3	8	60	0.019	0.30	0.95	0.98
Algorithm 4	**6**	**70**	**0.023**	**0.32**	**0.94**	**0.98**
Algorithm 5	5	75	0.023	0.33	0.94	0.97
Random	6	70	0.019	0.41	0.90	0.93

Note: Algorithm 1: SE(θ) ≤ 0.3; Algorithm 2: SE(θ) ≤ 0.3 OR Δθ < 0.005; Algorithm 3: SE(θ) ≤ 0.3 OR Δθ < 0.01; Algorithm 4: SE(θ) ≤ 0.3 OR Δθ < 0.05; Algorithm 5: SE(θ) ≤ 0.3 OR Δθ < 0.1 Algorithm selected is displayed in bold. One-dimensional model General Depression (20 items).

TABLE 3 Correlations between theta (θ) scores using the full test, and CAT Algorithm 4 with the direct scores (sums) of General Depression and the 18 facets of the IDAS-II and external variables.

IDAS-II Facets	Full test	CAT: Algorithm 4
General depression	0.94	0.92
Dysphoria	**0.93**	**0.91**
Lassitude	**0.74**	**0.72**
Insomnia	**0.70**	**0.68**
Suicidality	**0.57**	**0.56**
Appetite loss	**0.59**	**0.57**
Appetite gain	0.40	0.39
Ill-Temper	0.70	0.68
Mania	0.62	0.60
Panic	0.75	0.74
Traumatic intrusions	0.72	0.70
Traumatic avoidance	0.45	0.44
Social anxiety	0.66	0.64
Claustrophobia	0.46	0.45
Checking	0.47	0.46
Ordering	0.27	0.26
Cleaning	0.27	0.26
Well-being (recode)	**0.36**	**0.35**
Euphoria	0.20	0.20
External variables		
BAI (n = 620)	0.66	0.65
BDI-II (n = 620)	0.74	0.73
OCI-R Washing (n = 620)	0.29	0.28
OCI-R Cheking (n = 620)	0.34	0.33
OCI-R Ordering (n = 620)	0.31	0.30
OCI-R Neutralizing (n = 620)	0.29	0.28
OCI-R Hoarding (n = 620)	0.30	0.31
OCI-R Obsessing (n = 620)	0.51	0.50
OCI-R Total (n = 620)	0.47	0.46
PCL-C (n = 620)	0.56	0.54
PID-5-SF Negative affect (n = 1071)	0.58	0.58
PID-5-SF Detachment (n = 1071)	0.56	0.55
PID-5-SF Antagonism (n = 1071)	0.14	0.14
PID-5-SF Disinhibition (n = 1071)	0.51	0.51
PID-5-SF Psychoticism (n = 1071)	0.50	0.49
WHODAS 2.0 (n = 1401)	0.65	0.65

Note: In bold the facets that contribute with items in the CAT. Unidimensional General Depression model (20 items).

required in order to reach an acceptable measurement error value (see Supporting Information S3: Table S3).

3.3.2 | Validity evidence based on the relations with the IDAS-II facets

Table 3 shows the correlations between the summative scores of the GD scale and the 18 facets of the IDAS-II with the theta scores estimated from the model with all items (full test) and those estimated with the selected adaptive model (CAT = algorithm 4). The estimates provided by the adaptive algorithm presented correlations very similar to those of the estimates based on the 20 items (full test) for all scales. The highest correlations corresponded to GD and Dysphoria (r = 0.91–0.94). In general, CAT estimates maintained high and significant correlations with all facets that provide items to the measure, except Well-Being, with lower correlations to other facets.

3.3.3 | Validity evidence based on the relations with external measures

Table 3 shows the correlations between the estimates based on the application of the 20 items and the CAT estimates based on the chosen adaptive algorithm. Again, similar calculated correlations were observed between estimates based on all items (full test) and those calculated with the adaptive algorithm. The estimates correlated positively and with larger effect sizes with BDI (r = 0.74 and 0.73) and BAI (r = 0.66 and 0.65).

3.3.4 | Multidimensional (bifactor) model (99 items)

The adaptive algorithms used in the bifactor model are shown in Table 4. The first algorithm, based only on the value of the measurement error, provided low standard error values. However, we consider it to be an inefficient option given the high number of items involved. The algorithm that presented the best relation between precision and efficiency is algorithm 3 (Gen: $SE(\theta) \leq 0.3$ and Specifics: $SE(\theta) \leq 0.5$) OR ($\Delta\theta < 0.01$). With a reduction of 70% (approx.) in the number of items used, it resulted in standard error values of 0.30 for the general factor and less than 0.50 for each of the specific factors. Moreover, the scores estimated in this adaptive procedure showed high correlations with the 'true theta' and 'full test' estimates. On the other hand, these correlations were almost identical to those presented by the scores of algorithms 1 and 2, which provided better precision but involve a larger number of items. Similar to the unidimensional model, having few items with a position less than zero, the scores of people with true theta values less than zero were estimated with low precision, thus requiring a larger number of items to reach an acceptable level of measurement error (see Supporting Information S3: Table S3).

TABLE 4 Simulation studies comparing the different termination criteria with the application of the full test or a random selection of items.

	M items	% Reduction	SE(θ)				Correlations							
							Internalizing		Distress		Fear		Mood	
			Internalizing	Distress	Fear	Mood	True	Full	True	Full	True	Full	True	Full
Full	99	0	0.20	0.42	0.37	0.33	0.98	1.0	0.89	1.0	0.91	1.0	0.93	1.0
Algorithm 1	45	55	0.27	0.47	0.42	0.42	0.97	0.98	0.86	0.97	0.88	0.97	0.88	0.95
Algorithm 2	35	65	0.29	0.48	0.43	0.44	0.95	0.98	0.86	0.96	0.88	0.97	0.87	0.94
Algorithm 3	**30**	**70**	**0.30**	**0.48**	**0.44**	**0.46**	**0.95**	**0.97**	**0.85**	**0.96**	**0.88**	**0.96**	**0.86**	**0.93**
Algorithm 4	18	82	0.36	0.52	0.47	0.54	0.93	0.95	0.83	0.93	0.86	0.94	0.80	0.87
Algorithm 5	14	86	0.40	0.55	0.50	0.57	0.91	0.93	0.82	0.91	0.85	0.92	0.77	0.84
Random	30	70	0.32	0.65	0.60	0.55	0.95	0.96	0.72	0.81	0.76	0.84	0.78	0.85

Note: Algorithm 1: Gen: SE(θ) ≤ 0.3 and Specifics: SE(θ) ≤ 0.5; Algorithm 2: (Gen: SE(θ) ≤ 0.3 and Specifics: SE(θ) ≤ 0.5) OR $\Delta\theta < 0.005$; Algorithm 3: (Gen: SE(θ) ≤ 0.3 and Specifics: SE(θ) ≤ 0.5) OR $\Delta\theta < 0.01$; Algorithm 4: (Gen: SE(θ) ≤ 0.3 and Specifics: SE(θ) ≤ 0.5) OR $\Delta\theta < 0.05$; A Algorithm 5: (Gen: SE(θ) ≤ 0.3 and Specifics: SE(θ) ≤ 0.5) OR $\Delta\theta < 0.1$. Bifactorial model (99 items).

3.3.5 | Validity evidence based on the relations with the IDAS-II facets

Table 5 shows the correlations between the IDAS-II scale scores and the CAT estimates. The results showed very similar values for the chosen adaptive model as for the model including all items. In both cases, the estimated scores correlated as expected with the IDAS-II facets. The general factor estimates correlated positively and significantly with all IDAS scores. The values of the correlations ranged from 0.24 (for Well-Being) to 0.87 (for Dysphoria). Estimates for distress correlated mostly with the scales of Insomnia, Panic, Dysphoria and Traumatic Intrusions. Estimates on the specific factor of Fear/obsessions were closely related to the facets of Cleaning, Ordering or Claustrophobia. Finally, the estimates of Mood appeared associated with the two facets that integrate it (Well-Being and Euphoria).

3.3.6 | Validity evidence based on the relations with external measures

Table 5 shows the correlations between the estimates based on the application of the full test and CAT algorithm with external measures. The results revealed that the estimates of Algorithm 3 were very similar to those obtained with the complete test, with similar correlations of both estimates with other variables external to IDAS-II. In both cases, the estimated scores on the general factor correlated positively and significantly with all the variables analyzed, with values ranging from 0.25 for PID-5-SF Antagonism to 0.75 for BDI. Distress estimates correlated most strongly with anxiety (BAI), depression (BDI), PCL-C and WHODAS 2.0 Total scores. The Fear dimension was related to a greater extent, as expected, with OCI scores. Mood was related to depression, PID-5-SF Detachement and Hypomania.

4 | DISCUSSION

The present work for the first time shows that the adaptive version of IDAS-II (IDAS-CAT) is an interesting alternative for administrations in contexts where administration time is short. The IDAS-CAT could constitute a useful tool for both research and applied or clinical settings in which a comprehensive coverage of the internalizing spectrum is preferred. Specifically, the results obtained in the simulation process in a large mixed sample show that with a considerable reduction in the length of the IDAS-II and maintaining an adequate measurement error, the unidimensional CAT model allows to adequately represent the 20 items of the GD scale. Also, the bifactor model mirrors the 99 items of the IDAS-II grouped into three specific factors (distress, fear and mood) and one general factor (internalizing). In addition, the present simulation study provides evidence of the validity of the scores obtained with the two adaptive procedures. These results imply administration time can be substantially shortened without a reduction in reliability of validity.

TABLE 5 Correlations between theta (θ) scores using the full test, and CAT Algorithm 3 with the direct scores (sums) of General Depression and the 18 facets of the IDAS-II and external variables.

	Internalizing		Distress		Fear		Mood	
IDAS-II facets	Full test	CAT: Algorithm 3	Full test	CAT: Algorithm 3	Full test	CAT: Algorithm 3	Full test	CAT: Algorithm 3
General depression	0.85	0.83	0.38	0.41	−0.05	−0.04	0.28	0.26
Dysphoria	0.87	0.85	0.28	0.31	−0.04	−0.03	0.20	0.19
Lassitude	0.74	0.72	0.17	0.21	−0.04	−0.03	0.07	0.08
Insomnia	0.60	0.61	0.72	0.71	0.05	0.04	0.09	0.08
Suicidality	0.61	0.58	0.07	0.09	−0.01	−0.01	0.14	0.14
Appetite loss	0.52	0.51	0.33	0.34	0.03	0.02	0.08	0.08
Appetite gain	0.48	0.46	0.01	0.03	−0.01	−0.01	−0.04	−0.02
Ill-Temper	0.75	0.73	0.18	0.20	−0.02	−0.02	0.04	0.05
Mania	0.75	0.73	0.08	0.11	0.13	0.13	−0.13	−0.11
Panic	0.78	0.77	0.27	0.29	0.01	0.02	0.11	0.11
Traumatic intrusions	0.74	0.72	0.22	0.25	−0.02	−0.02	0.10	0.10
Traumatic avoidance	0.57	0.56	0.01	0.03	0.19	0.19	−0.17	−0.15
Social anxiety	0.74	0.72	0.07	0.10	0.06	0.06	0.09	0.09
Claustrophobia	0.56	0.55	0.06	0.07	0.44	0.43	−0.02	−0.02
Checking	0.60	0.58	0.01	0.03	0.31	0.29	−0.17	−0.16
Ordering	0.43	0.41	−0.05	−0.03	0.42	0.41	−0.28	−0.26
Cleaning	0.34	0.33	0.05	0.06	0.79	0.77	−0.12	−0.11
Well-being (recode)	0.24	0.24	0.15	0.16	−0.15	−0.14	0.87	0.81
Euphoria	0.41	0.40	−0.10	−0.08	0.18	0.17	−0.60	−0.55
External variables								
BAI (n = 620)	0.72	0.70	0.26	0.29	0.22	0.21	0.12	0.11
BDI-II (n = 620)	0.75	0.73	0.28	0.30	0.07	0.07	0.29	0.28
OCI-R Washing (n = 620)	0.37	0.35	0.05	0.06	0.54	0.51	−0.06	−0.06
OCI-R Checking (n = 620)	0.43	0.40	0.11	0.12	0.36	0.34	−0.04	−0.06
OCI-R Ordering (n = 620)	0.38	0.38	0.09	0.10	0.35	0.33	−0.07	−0.08
OCI-R Neutralizing (n = 620)	0.39	0.36	0.02	0.04	0.32	0.30	−0.10	−0.10
OCI-R Hoarding (n = 620)	0.36	0.35	0.04	0.06	0.22	0.22	−0.02	−0.02
OCI-R Obsessing (n = 620)	0.58	0.56	0.11	0.14	0.16	0.14	0.05	0.06
OCI-R Total (n = 620)	0.58	0.55	0.08	0.11	0.44	0.41	−0.05	−0.05
PCL-C (n = 620)	0.59	0.56	0.19	0.22	0.07	0.08	0.09	0.10
PID-5-SF Neagative affect (n = 1071)	0.63	0.61	0.11	0.13	0.07	0.07	0.10	0.09
PID-5-SF Detachment (n = 1071)	0.58	0.57	0.08	0.10	−0.04	−0.02	0.28	0.27
PID-5-SF Antagonism (n = 1071)	0.25	0.23	−0.10	−0.08	0.01	0.02	−0.12	−0.09
PID-5-SF Disinhibition (n = 1071)	0.59	0.57	0.01	0.04	−0.06	−0.05	0.08	0.09
PID-5-SF Psychoticism (n = 1071)	0.60	0.58	0.02	0.05	−0.01	0.02	0.04	0.05
WHODAS 2.0 (n = 1401)	0.66	0.65	0.17	0.19	0.01	0.01	0.17	0.15

Note: Bifactorial model (99 ítems).

Firstly, the CFAs conducted provide evidence of validity based on the internal structure of the test of IDAS-II. Our data support the factor structure proposed by Watson et al. (2012; see also in other languages: Irak & Albayrak, 2020; Wester et al., 2022): the 20 DG items fit a unidimensional model, while the 99 full-test items allow replicating the three-factor structure (distress, fear and mood). Furthermore, the observation of an adequate fit of the bifactor model is further evidence in favor of the internalizing spectrum defined in the HiTOP (Kotov et al., 2017; Waszczuk et al., 2017). Similar results are observed in the work of Sunderland et al. (2019), with one general internalizing factor and five specific factors using items from different instruments.

Secondly, the results of the present work show that the simulated versions of the IDAS-CAT improve the metric properties of the corresponding non-adaptive full test versions. In particular, the scores generated with the adaptive algorithms have a very low average measurement error with significantly fewer items. Maintaining a high accuracy (SE(θ) close to 0.30 for GD and internalizing general factor and SE(θ) below 0.50 for the specific factors) relevant levels of efficiency are obtained. This results into a reduction of 70% of items in both cases. Similar values are obtained in simulations conducted in previous studies with the 21-item of BDI (71% reduction; Gardner et al., 2004), in a one-dimensional model with the 28-item of the PROMIS depressive symptoms bank (72% reduction with SE(θ) = 0.26; Choi et al., 2010) or with the 20-item unidimensional model of externalizing disinhibition (70% reduction with an SE(θ) of 0.34 under the assumption of normal distribution; Sunderland et al., 2017). As for the bifactor model, our data are comparable with those obtained by Sunderland et al. (2019) (67% reduction for SE (θ) = 0.27 on the general factor and SE(θ) values of 0.66, 0.50, 0.45, 0.62 and 0.48 for the specific factors). As noted by several authors, scores on the specific factors are always measured less accurately than those on the general factor (Sahin & Gelbal, 2020; Seo & Weiss, 2015). Moreover, the correlations of the sub-factors with their respective true theta are usually lower. One possible reason may be the very nature of the bifactor model, which requires factor loadings on the general factor to be higher than those on the specific factors (Reise et al., 2007).

Finally, the results show high correlations of the scores estimated with the adaptive algorithms and the estimates based on the full test, as well as correlations with external criteria almost equal to those generated with the full test. The relationships found are similar to those found in previous works (De la Rosa et al., 2020; Irak & Albayrak, 2020; Sunderland et al., 2019; Watson et al., 2012; Wester et al., 2022): both the general factor (internalizing) and GD are closely related to scores for depression, anxiety, PTSD or OCD.

Overall, the results presented in this paper have additional advantages over previous CAT simulation works in that they use items from the same instrument. This allows the items included in both adaptive procedures to have the same response format and the same time frame. In addition, for the generation of the bifactor algorithm, the content of the items has been considered for the CAT application process, this way ensuring proportionality of the dimension including in the three subfactors (Chalmers, 2016). This allows the bifactor adaptive version of the IDAS-II to represent an effective and representative measure of the internalizing spectrum with its various subfactors.

Although the results presented here constitute a significant contribution to the scientific literature, there are a number of limitations that need to be considered. First, it should be noted that these data are simulated, an thus not comprised with real applications. Although research has shown that CAT simulations are quite similar and consistent with real applications (Kocalevent et al., 2009), we are aware that the 'real' responses of participants may be affected by different factors such as the environment, personal situation or other possibly influencing aspects. Similarly, it would be interesting for further studies to see to what extent the likely presence of mismatched response patterns may affect the accuracy and efficiency of IDAS-CAT.

Second, comparable to CAT procedures on other variables (see, for example, Choi et al., 2010; Peersmann et al., 2022; Sunderland et al., 2019; Tsaouis et al., 2021), in our study, CAT provides greater precision and efficiency for the core values of ability level. Thus, individuals measured with lower precision and requiring a higher number of items tend to be those whom correspond to more extreme values along the dimensional construct (being either well below or well above the norm in terms of standard deviations). In our case, this is especially true for people with low levels of internalizing or GD. These considerations should be taken into account by future studies with real applications of the CAT.

Future work could also be extended to explore both efficiency (in terms of the number of items needed) and precision (in terms of the measurement error achieved) when needing to place people above or below a cut-off point. In this case, the selection of items would focus on choosing the most discriminative items around that cut-off point. This may predictably require a greater number of items and/or less precision, especially for individuals whose values are far from the cut-off point.

5 | CONCLUSIONS

Despite the limitations mentioned above, this work offers, through the IDAS-CAT simulation, a useful, reliable and fast tool for measuring the variables and dimensions included in the IDAS-II. It can provide a fast and efficient evaluation that takes comorbidity of internalizing disorders into account by measuring broad and specific levels of psychopathology. Consequently, it can provide mental health clinicians a comprehensive and efficient measure of the internalizing spectrum. Moreover, it can also be used to screen for individuals at risk for internalizing pathology in general the general community and this way identify individuals in need of care. Independent evaluations, with different samples and in real contexts, are now needed to provide further information on its ecological validity and corroborate the usefulness of a computerized adaptive version of IDAS-II.

AUTHOR CONTRIBUTIONS

M. Sanchez-Garcia: Conceptualization; methodology; formal analysis; visualization; writing - original draft; writing - review & editing. **A. De la Rosa-Caceres**: Methodology; software; data curation; writing - review & editing; investigation. **G. Rossi**: Writing - review & editing; validation; writing - original draft; supervision. **C. Diaz-Batanero**: Conceptualization; methodology; writing - original draft; funding acquisition; resources; supervision; project administration.

ACKNOWLEDGMENTS

This work was supported by the grant "Reliable and clinical relevant change of Inventory of Depression and Anxiety Symptoms II – IDAS-II: a longitudinal clinical utility study (RELY-IDAS-II)", project PID2020-116187RB-I00 on Proyectos I+D+i 2020 "Retos del Conocimiento" funded by Ministerio de Ciencia e Innovación (Spain) (MICIU/AEI/10.13039/501100011033) and grant number FPU19/00144 funded by Ministerio de Universidades (Spain) (MICIU/AEI/10.13039/501100011033) and by ESF Investing in your future.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest to report.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

ETHICS STATEMENT

The study was approved by the Bioethics Committee of the Province of Huelva (Junta de Andalucía, Spain) (No. PY18-4584).

ORCID

M. Sanchez-Garcia https://orcid.org/0000-0003-3375-8347
A. De la Rosa-Cáceres https://orcid.org/0000-0003-3719-0840
G. Rossi https://orcid.org/0000-0003-1062-0070
C. Diaz-Batanero https://orcid.org/0000-0003-3392-4683

REFERENCES

Achenbach, T. M., Ivanova, M. Y., Rescorla, L. A., Turner, L. V., & Althoff, R. R. (2016). Internalizing/externalizing problems: Review and recommendations for clinical and research applications. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(8), 647–656. https://doi.org/10.1016/j.jaac.2016.05.012

Barlow, D. H., Allen, L. B., & Choate, M. L. (2016). Toward a unified treatment for emotional disorders. *Behavior Therapy*, *47*(6), 838–853. https://doi.org/10.1016/j.beth.2016.11.005

Barlow, D. H., Farchione, T. J., Bullis, J. R., Gallagher, M. W., Murray-Latin, H., Sauer-Zavala, S., Bentley, K. H., Thompson-Hollands, J., Conklin, L. R., Boswell, J. F., Ametaj, A., Carl, J. R., Boettcher, H. T., & Cassiello-Robbins, C. (2017). The unified protocol for transdiagnostic treatment of emotional disorders compared with diagnosis-specific protocols for anxiety disorders: A randomized clinical trial. *JAMA Psychiatry*, *74*(9), 875–884. https://doi.org/10.1001/jamapsychiatry.2017.2164

Barlow, D. H., Farchione, T. J., Fairholme, C., Ellard, K. K., Boisseau, C., Allen, L., & Ehrenreich-May, J. (2011). *Unified protocol for the transdiagnostic treatment of emotional disorders: Therapist guide*. Oxford University Press.

Barnard, J. J. (2018). From simulation to implementation: Two CAT case studies. *Practical Assessment, Research and Evaluation*, *23*, Article14.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for Beck depression inventory-II*. Psychological Corp.

Beck, A. T. y, & Steer, R. (1993). *Beck anxiety inventory manual*. Psychological Corporation.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–39. https://doi.org/10.18637/jss.v071.i05

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research: an international journal of quality of life aspects of treatment, care and rehabilitation*, *19*(1), 125–136. https://doi.org/10.1007/s11136-009-9560-5

De la Rosa-Cáceres, A., López, J. R., Fernández Calderón, F., Lozano-Rojas, O., Moraleda-Barreno, E., & Díaz-Batanero, C. (2020). Discriminative capacity for functional impairment of the Personality Inventory for DSM-5 Short Form in patients with substance use disorder. *Adicciones*, *32*(4), 307–309. https://doi.org/10.20882/adicciones.1357

De la Rosa-Caceres, A., Lozano, O. M., Sanchez-Garcia, M., Fernandez-Calderon, F., Rossi, G., & Diaz-Batanero, C. (2023). Assessing internalizing symptoms and their relation with levels of impairment: Evidence-based Cutoffs for interpreting inventory of depression and anxiety symptoms (IDAS-II) scores. *Journal of Psychopathology and Behavioral Assessment*, *45*(1), 170–180. https://doi.org/10.1007/s10862-022-10008-6

De la Rosa-Caceres, A., Stasik-O'Brien, S. M., Rojas, A. J., Sanchez-Garcia, M., Lozano, O. M., & Diaz-Batanero, C. (2020). Spanish adaptation of the Inventory of Depression and Anxiety Symptoms (IDAS-II) and a study of its psychometric properties. *Journal of Affective Disorders*, *271*, 81–90. https://doi.org/10.1016/j.jad.2020.03.187

Díaz-Batanero, C., Ramírez-López, J., Domínguez-Salas, S., Fernández-Calderón, F., & Lozano, Ó. M. (2019). Personality Inventory for DSM-5–Short Form (PID-5-SF): Reliability, factorial structure, and relationship with functional impairment in dual diagnosis patients. *Assessment*, *26*(5), 853–866. https://doi.org/10.1177/1073191117739980

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Ferrando, P. J., Lorenzo-Seva, U., Hernández-Dorado, A., & Muñiz, J. (2022). Decálogo para el Análisis Factorial de los Ítems de un Test. *Psicothema*, *34*(1), 7–17. [Decalogue for the factor analysis of test items]. https://doi.org/10.7334/psicothema2021.456

Feuerstahler, L. M., Waller, N., & MacDonald, A., III. (2020). Improving measurement precision in Experimental psychopathology using item response theory. *Educational and Psychological Measurement*, *80*(4), 695–725. https://doi.org/10.1177/0013164419892049

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment*, *14*(4), 485–496. https://doi.org/10.1037/1040-3590.14.4.485

Fullana, M. A., Tortella-Feliu, M., Caseras, X., Andión, Ó., Torrubia, R., & Mataix-Cols, D. (2005). Psychometric properties of the Spanish version of the Obsessive-Compulsive Inventory-Revised in a

non-clinical sample. *Journal of Anxiety Disorders*, *19*(8), 893–903. https://doi.org/10.1016/j.janxdis.2004.10.004

Gardner, W., Shear, K., Kelleher, K. J., Pajer, K. A., Mammen, O., Buysse, D., & Frank, E. (2004). Computerized adaptive measurement of depression: A simulation study. *BMC Psychiatry*, *4*(1), 13. https://doi.org/10.1186/1471-244X-4-13

Gibbons, R., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19. https://doi.org/10.1177/0146621606289485

Gibbons, R., Weiss, D., Pilkonis, P., Frank, E., Moore, T., Kim, J., & Kupfer, D. (2014). Development of the catanx: A computerized adaptive test for anxiety. *American Journal of Psychiatry*, *171*(2), 187–194. https://doi.org/10.1176/appi.ajp.2013.13020178

Gibbons, R., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology*, *12*(1), 83–104. https://doi.org/10.1146/annurev-clinpsy-021815-093634

Gibbons, R., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). The CAT-DI: Development of a computerized adaptive test for depression. *Archives of General Psychiatry*, *69*(11), 1104–1112. https://doi.org/10.1001/archgenpsychiatry.2012.14

Guinart, D., de Filippis, R., Rosson, S., Patil, B., Prizgint, L., Talasazan, N., Meltzer, H., Kane, J. M., & Gibbons, R. D. (2021). Development and validation of a computerized adaptive assessment tool for discrimination and measurement of psychotic symptoms. *Schizophrenia Bulletin*, *47*(3), 644–652. https://doi.org/10.1093/schbul/sbaa168

Han, K. T., & Kosinski, M. (2014). Software tools for multistage testing simulations. In D. Yan, A. A. von-Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 411–420). CRC Press: Taylor&Francis Group.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hulvershorn, L. A., Adams, Z. W., Smoker, M. P., Aalsma, M. C., & Gibbons, R. D. (2022). Development of a computerized adaptive substance use disorder scale for screening, measurement and diagnosis–The CAT-SUD-E. *Drug and Alcohol Dependence Reports*, *3*, 100047. https://doi.org/10.1016/j.dadr.2022.100047

Irak, M., & Albayrak, E. O. (2020). Psychometric properties of the expanded version of the inventory of depression and anxiety symptoms in a Turkish population. *Psychological Reports*, *123*(2), 517–545. https://doi.org/10.1177/0033294118813844

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*, 109. https://doi.org/10.3389/fpsyg.2016.00109

Kang, T., & Chen, T. T. (2008). Performance of the generalized S- X2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kessler, R. C., Sampson, N. A., Berglund, P., Gruber, M. J., Al-Hamzawi, A., Andrade, L., Bunting, B., Demyttenaere, K., Florescu, S., De Girolamo, G., Gureje, O., He, Y., Hu, C., Huang, Y., Karam, E., Kovess-Masfety, V., Lee, S., Levinson, D., Medina Mora, M. E. & Wilcox, M. A. (2015). Anxious and non-anxious major depressive disorder in the World Health Organization world mental health surveys. *Epidemiology and Psychiatric Sciences*, *24*(3), 210–226. https://doi.org/10.1017/S2045796015000189

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., Kleiber, D., & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*, *62*(3), 278–287. https://doi.org/10.1016/j.jclinepi.2008.03.003

Kotov, R., Waszczuk, M. A., Krueger, R. F., Forbes, M. K., Watson, D., Clark, L. A., Achenbach, T. M., Althoff, R. R., Ivanova, M. Y., Michael Bagby, R., Brown, T. A., Carpenter, W. T., Caspi, A., Moffitt, T. E., Eaton, N. R., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E. & Zimmerman, M. (2017). The hierarchical Taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological Medicine*, *42*(9), 1879–1890. https://doi.org/10.1017/s0033291711002674

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology*, *2*(1), 111–133. https://doi.org/10.1146/annurev.clinpsy.2.022305.095213

Lamers, F., van Oppen, P., Comijs, H. C., Smit, J. H., Spinhoven, P., van Balkom, A. J., Nolen, W. A., Zitman, F. G., Beekman, A. T., & Penninx, B. W. (2011). Comorbidity patterns of anxiety and depressive disorders in a large cohort study: The Netherlands study of depression and anxiety (NESDA). *Journal of Clinical Psychiatry*, *71*(03), 341–348. https://doi.org/10.4088/JCP.10m06176blu

Luecht, R. M., de Champlain, A., & Nungester, R. J. (1998). Maintaining content validity in computerized adaptive testing. *Advances in Health Sciences Education*, *3*(1), 29–41. https://doi.org/10.1023/A:1009789314011

Maples, J. L., Carter, N., Few, L. R., Crego, C., Gore, W. L., Samuel, D. B., Williamson, R. L., Lynam, D. R., Widiger, T. A., Markon, K. E., Krueger, R. F., & Miller, J. D. (2015). Testing whether the DSM-5 personality disorder trait model can be measured within a reduced set of items: An item response theory investigation of the Personality Inventory for DSM-5. *Psychological Assessment*, *27*(4), 1195–1210. https://doi.org/10.1037/pas0000120

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. W eiss (Ed.), *New horizons in testing* (pp. 223–236). Academic Press.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, *23*(3), 187–194. https://doi.org/10.1177/01466219922031310

Mindrila, D. (2010). Maximum likelihood (ML) and diagonally weighted Least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society (IJDS)*, *1*(1), 60–66. https://doi.org/10.20533/ijds.2040.2570.2010.0010

Mustanski, B., Whitton, S. W., Newcomb, M. E., Clifford, A., Ryan, D. T., & Gibbons, R. D. (2021). Predicting suicidality using a computer adaptive test: Two longitudinal studies of sexual and gender minority youth. *Journal of Consulting and Clinical Psychology*, *89*(3), 166–175. https://doi.org/10.1037/ccp0000531

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD Checklist: Detection and evaluation of impact. *Psychological Assessment*, *14*(1), 50–59. https://doi.org/10.1037/1040-3590.14.1.50

Peersmann, S. H. M., Luijten, M. A. J., Haverman, L., Terwee, C. B., Grootenhuis, M. A. & Raphaële, R. L. (2022). Psychometric properties and CAT performance of the PROMIS pediatric sleep disturbance, sleep-related impairment, and fatigue item banks in Dutch children and adolescents. *Psychological Assessment*, *34*(9), 860–869. https://doi.org/10.1037/pas0001150

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures.

Quality of Life Research, *16*(S1), 19–31. https://doi.org/10.1007/s11136-007-9183-7

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sahin, M. D., & Gelbal, S. (2020). Development of a multidimensional computerized adaptive test based on the bifactor model. *International Journal of Assessment Tools in Education*, *7*(3), 323–342. https://doi.org/10.21449/ijate.707199

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer. https://doi.org/10.1007/978-1-4757-2691-6_5

Sanz, J., & Navarro, M. E. (2003). Propiedades psicométricas de una versión española del inventario de ansiedad de beck (BAI) en estudiantes universitarios. [The psychometric properties of a spanish version of the Beck Anxiety Inventory (BAI) in a university students sample] *Ansiedad y Estrés*. https://www.proquest.com/scholarly-journals/propiedades-psicométricas-de-una-versión-española/docview/620297630/se-2

Sanz, J., Navarro, M. E., & Vázquez, C. (2003). Adaptación española del inventario para la depresión de beck--II(BDI-II): 1. Propiedades psicométricas en estudiantes universitarios. [Spanish adaptation of the Beck Depression Inventory-II (BDI-II): 1. Psychometric properties with university students] *Análisis y Modificación De Conducta*. https://www.proquest.com/scholarly-journals/adaptación-española-del-inventario-para-la/docview/620382583/se-2

Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement*, *75*(6), 954–978. https://doi.org/10.1177/0013164415575147

Stochl, J., Böhnke, J. R., Pickett, K. E., & Croudace, T. J. (2016). Computerized adaptive testing of population psychological distress: Simulation-based evaluation of GHQ-30. *Social Psychiatry and Psychiatric Epidemiology*, *51*(6), 895–906. https://doi.org/10.1007/s00127-015-1157-4

Sunderland, M., Batterham, P., Carragher, N., Calear, A., & Slade, T. (2019). Developing and validating a computerized adaptive test to measure broad and specific factors of internalizing in a community sample. *Assessment*, *26*(6), 1030–1045. https://doi.org/10.1177/1073191117707817

Sunderland, M., Slade, T., Krueger, R. F., Markon, K. E., Patrick, C. J., & Kramer, M. D. (2017). Efficiently measuring dimensions of the externalizing spectrum model: Development of the externalizing spectrum inventory-computerized adaptive test (ESI-CAT). *Psychological Assessment*, *29*(7), 868–880. https://doi.org/10.1037/pas0000384

Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, *60*, 41–63. https://doi.org/10.1016/j.jsp.2016.11.001

Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2021). Evaluating a computerized adaptive testing version of a cognitive ability test using a simulation study. *Journal of Psychoeducational Assessment*, *39*(8), 954–968. https://doi.org/10.1177/07342829211027753

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & G. A. Glas (Eds.), *Computerized adaptive testing: Theory and practice*. Springer.

Vazquez-Barquero, J. L., Vazquez Bourgon, E., Herrera Castanedo, S., Saiz, J., Uriarte, M., Morales, F., Gaite, L., Herran, A., & Ustun, T. B. (2000). Spanish version of the new world health organization disability assessment Schedule II (WHO-DAS-II): Initial phase of development and pilot study. *Actas Espanolas de Psiquiatria*, *28*(2), 77–87. PMID: 10937388.

Vittengl, J. R., Jarrett, R. B., Ro, E., & Clark, L. A. (2023). How can the DSM-5 alternative model of personality disorders advance understanding of depression? *Journal of Affective Disorders*, *320*, 254–262. https://doi.org/10.1016/j.jad.2022.09.146

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adapative testing: A primer* (2nd ed., pp. 1–22). Lawrence Erlbaum.

Waszczuk, M. A., Kotov, R., Ruggero, C., Gamez, W., & Watson, D. (2017). Hierarchical structure of emotional disorders: From individual symptoms to the spectrum. *Journal of Abnormal Psychology*, *126*(5), 613–634. https://doi.org/10.1037/abn0000264

Watson, D., O'Hara, M. W., Naragon-Gainey, K., Koffel, E., Chmielewski, M., Kotov, R., Stasik, S. M., & Ruggero, C. J. (2012). Development and validation of new anxiety and bipolar symptom scales for an expanded version of the IDAS (the IDAS-II). *Assessment*, *19*(4), 399–420. https://doi.org/10.1177/1073191112449857

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A., & Keane, T. M. (1993). The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility. In *Paper presented at the annual meeting of the International Society for traumatic stress studies*. Bearman.

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods*, *18*(1), 53–70. https://doi.org/10.1037/a0031607

Wester, R. A., Rubel, J., Zimmermann, J., Hall, M., Kaven, L., & Watson, D. (2022). Development and validation of the inventory of depression and anxiety symptoms–II—German version. *Psychological Assessment*, *34*(12), e88–e99. https://doi.org/10.1037/pas0001185

World Health Organization. (2000). *Disability assessment Schedule II (WHO-DAS II)*. WHO.

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment*, *15*(4), 478–495. https://doi.org/10.1037/1040-3590.15.4.478

Yi, Q., & Chang, H. H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 359–378. https://doi.org/10.1348/000711003770480084

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sanchez-Garcia, M., De la Rosa-Cáceres, A., Rossi, G., & Diaz-Batanero, C. (2024). Developing an accurate and efficient tool for the internalizing spectrum: A simulation study of the adaptive algorithm to the Inventory of Depression and Anxiety Symptoms II (IDAS-II). *International Journal of Methods in Psychiatric Research*, e2032. https://doi.org/10.1002/mpr.2032