jove

Cross-cultural Adaptation and Psychometric Validation of a Structured Interview for Psychiatric Assessment

Sílvia Almeida[1,2], **Pedro Castro-Rodrigues**[1,3,4], **J. Bernardo Barahona-Corrêa**[1,3], **Telmo Mourinho Baptista**[5], **Jaime Grácio**[1,3], **Albino J. Oliveira-Maia**[1,3]

[1] Champalimaud Research and Clinical Centre, Champalimaud Foundation [2] Graduate Programme in Clinical and Health Psychology, Faculdade de Psicologia da Universidade de Lisboa [3] NOVA Medical School, Faculdade de Ciências Médicas, NMS, FCM, Universidade NOVA de Lisboa [4] Centro Hospitalar Psiquiátrico de Lisboa [5] Faculdade de Psicologia da Universidade de Lisboa

Corresponding Authors

Sílvia Almeida
silvia.almeida@research.fchampalimaud.org

Albino J. Oliveira-Maia
albino.maia@neuro.fchampalimaud.org

Citation

Almeida, S., Castro-Rodrigues, P., Barahona-Corrêa, J.B., Baptista, T.M., Grácio, J., Oliveira-Maia, A.J. Cross-cultural Adaptation and Psychometric Validation of a Structured Interview for Psychiatric Assessment. *J. Vis. Exp.* (224), e68710, doi:10.3791/68710 (2025).

Date Published

October 31, 2025

DOI

10.3791/68710

URL

jove.com/video/68710

Abstract

Psychiatric disorders are a significant cause of long-term disability and mortality. Although treatment is available, diagnostic accuracy is critical to provide adequate evidence-based treatment and to develop novel therapies. To inform the diagnostic process during clinical interviews, the use of validated assessment measures, including self-report questionnaires and structured interviews, is highly recommended. However, such instruments must have excellent psychometric properties, particularly regarding reliability and validity, to ensure accurate and interpretable data for each individual. Furthermore, applying an instrument in a new country, context, or language requires a formal cultural adaptation. This process is mandatory to ensure that the findings from the adapted version are equivalent to those of the original questionnaire.

Here, we describe a detailed protocol for cultural adaptation and comprehensive psychometric validation of a psychometric instrument. Specifically, we outline the steps for selecting the measure, conducting the experimental procedures, and performing statistical analyses required to establish the instrument's psychometric properties, including reliability, construct validity, and criterion validity for diagnosis of a psychiatric disorder. Our primary purpose is to present a transparent, standardized method for culturally adapting and validating psychometric instruments. Such a procedure helps minimize confounding factors and undesired variability in future applications and research. We expect that this protocol, including a range of empirically supported methods, will be useful in research settings for the cultural adaptation of psychometric instruments for psychiatric assessment.

Introduction

The World Health Organization (WHO) defines psychiatric disorders as a combination of abnormal thoughts, perceptions, emotions, behavior, and interpersonal relationships[1]. These conditions represent a significant cause of long-term disability and mortality[2]. The broad spectrum of these disorders includes major depressive disorder, obsessive-compulsive disorder (OCD), generalized anxiety disorder, and bipolar disorder[3]. Although there are treatment options for each of these mental disorders, the accuracy of the diagnosis is critical to provide adequate evidence-based treatment[4].

Regarding diagnostic assessment, several guidelines, such as those from the National Institute for Health & Clinical Excellence, recommend the use of validated assessment measures relevant to the disorder that is being assessed[5], in order to provide additional information for the clinician[6]. There are several instruments for a variety of mental disorders[7], developed to screen, diagnose, and assess symptom severity or response to treatment[8,9]. However, before being considered adequate, an instrument must offer accurate, valid, and interpretable data for the population to be assessed[10]. Importantly, the quality of the information about a specific individual depends on the psychometric properties of the instrument used[11]. To reduce bias in the testing process, from application to interpretation of the results, psychological measures should be standardized[8]. This was the main reason for the creation of the *Standards for Educational and Psychological Testing*, as a basis for evaluating tests, testing practices, and the impact of test use[12]. Equally important is the fact that most instruments were developed in English-speaking countries[13] making cultural and linguistic adaptation necessary prior to use in a new country, culture, and/or language, to reach equivalence between the original (source) and the newly adapted (target) versions of the questionnaire[14].

When an established instrument is not available in a specific language or culture, researchers face a choice between two main strategies: developing a new, context-specific instrument or performing a cross-cultural adaptation of an existing, well-validated measure[15]. While the development of a novel instrument can ensure maximum cultural specificity, it is an extremely resource- and time-intensive process that may take years[16]. In contrast, the adaptation of an established 'gold-standard' instrument offers distinct advantages. This approach is often more efficient and, critically, it allows for the cross-cultural comparison of findings from different populations, which is a primary goal of adapting measures rather than creating new ones[17].

The International Test Commission has developed guidelines for cross-cultural translation and adaptation of psychological instruments[17]. Translation can be considered the first stage of the adaptation process[18], and can be conducted using one or both of the two most popular methods of test translation: (a) translation and back-translation, or (b) two independent translations that are compared by a third person[19]. The cultural adaptation process requires that, in addition to an exact translation, an adaptation process be conducted to maximize semantic, idiomatic, experiential, and conceptual equivalence between the original measure and those that are developed from it[14,20]. Finally, the psychometric properties of a translated instrument should be evaluated in order to compare them with the original measure in the primary language[20]. Specifically, it is important to assess reliability and validity[8,9,21], assuring, respectively, that the instrument results in a consistent measurement, and that it measures the intended construct[22].

Reliability refers to the reproducibility of a test result when obtained at different times, in different settings, or by different interviewers, regarding coherence, stability, equivalence, and homogeneity[23,24,25]. It can be evaluated through several methods, including assessments of test-retest, alternate forms, split-half reliability, as well as internal consistency[8,22,26], determining whether the measures are sufficiently consistent and free from measurement error[8]. Although an instrument that is not reliable cannot be valid, a reliable instrument can sometimes be invalid[10]. Validity is considered according to three categories[27,28], namely content validity, construct validity, and criterion validity. The concept of content validity concerns the extent to which a test adequately samples the dimension it is intended to measure[22], while construct validity, including convergent and discriminant validity (sometimes referred to as divergent validity[29]), represents the degree to which the variance of the measure is linked with the variance of the underlying construct[30,31]. Criterion validity is based on relationships between test scores[9] and should be assessed using another measure of the same construct, ideally a widely accepted measure that is considered the gold standard[8,28]. This category of validity is especially important to understand whether a measure can be used to make predictions and/or decisions about patients[25], which is the case in establishing a diagnosis.

Numerous guidelines for the cross-cultural adaptation of psychometric instruments have been published to aid researchers in this complex process[17,32]. However, systematic reviews of this literature have highlighted a lack of a single, unified consensus on the best methodology to follow[33]. Furthermore, many existing guides, while valuable, may focus more on the initial linguistic translation than on the equally critical subsequent psychometric validation required to ensure an instrument is ethically sound for clinical use[19]. This creates a need for a detailed, replicable protocol that integrates both the adaptation and a comprehensive validation phase into a single, step-by-step framework.

Standardized research practices focusing on the validation of psychometric measures are thus essential. The method described in this paper will provide researchers and clinicians with a detailed protocol to perform cultural adaptation of a psychometric measure and, specifically, to assess criterion validity for the diagnosis of a psychiatric disorder. To help readers assess its applicability and to ensure replicability, the protocol includes key practical details, such as sample size considerations, the rationale for multi-session administration timings, and a discussion of known limitations. For that purpose, we will use, as an example, the validation study of the European Portuguese Yale-Brown Obsessive-Compulsive Scale-Second Edition (PY-BOCS-II)[34], in which a similar protocol was used to clarify the factor structure and criterion validity of the PY-BOCS-II for the diagnosis of OCD in adults. Therefore, this protocol can also be used for future validation studies of Y-BOCS-II in other contexts or languages.

Protocol

The procedures described here were developed to collect the data described by Castro-Rodrigues et al.[34]. The protocol was prepared in accordance with the Declaration of Helsinki, and participants were informed of the possibility of withdrawing from the study at any time. It was reviewed and approved by the Ethics Committees of the Champalimaud Foundation (approval granted on October 22, 2014) and Centro Hospitalar Psiquiátrico de Lisboa (approval granted on November 14, 2014). Use of this protocol for other projects or in other locations should be performed only after approval by

local Ethics Committees and/or other competent authorities at that location. Specific examples regarding the Portuguese adaptation of the Y-BOCS-II[34] are given to illustrate some of the steps, and specific instructions for the validation of the Y-BOCS-II for other languages/contexts are provided.

1. Selection of the scale of interest

1. Define the construct and diagnosis of interest. First, clearly articulate the specific clinical construct to be assessed. This may be a broad diagnostic category (e.g., OCD) or a specific dimension within it (e.g., symptom severity). This definition will guide the entire adaptation and validation process.

   NOTE: Castro-Rodrigues et al.[34] were interested in symptom severity and diagnosis of OCD.

2. Select a measure with adequate psychometric properties to study the construct of interest, ideally after conducting a review of the literature, not only to define the measure but also to confirm that the planned work is not redundant with prior publications. For example, as in the study by Castro-Rodrigues et al.[34], to assess OCD, select the Y-BOCS-II[35], as it is internationally recognized as the gold standard for this purpose.

   NOTE: The Y-BOCS-II[35], a clinician-administered interview for adults, allows for a detailed assessment of OCD. The instrument is composed of two main parts: a 67 item Symptom Checklist to identify and classify current and past obsessions, compulsions, and avoidance behaviors, and a 10 item Severity Scale to rate the severity of those symptoms.

3. Obtain permission to adapt the scale. Identify the authors and the copyright holder of the original instrument. Contact them to formally request permission for the translation and cultural adaptation of the instrument for research purposes. As was done for the Y-BOCS-II[34], ensure this permission is granted before beginning the translation process.

2. Selection of other measures for assessment of psychometric properties of the scale

1. Select a measure for criterion validity. To establish criterion validity for diagnosis, choose a measure that is considered the gold standard (if available) to discriminate between participants with and without the diagnosis of interest. To follow the example of Castro-Rodrigues et al.[34], use the OCD subscale of the Structured Clinical Interview for the DSM-IV Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) (SCID-OCD)[36,37], for this purpose.

   NOTE: The OCD subscale of the Structured Clinical Interview for the DSM-IV Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) (SCID-OCD)[36,37], is a semi-structured interview allowing for the diagnosis of current OCD, according to DSM-IV criteria.

2. Select a measure to assess comorbidity and to identify the presence of psychiatric exclusion criteria (see details below). Choose a structured psychiatric interview that covers a range of diagnoses. To follow the previous example[34], use the Mini-International Neuropsychiatric Interview (MINI)[38], a brief structured clinical interview based on rapid screening of DSM-IV diagnostic criteria.

   NOTE: Divided into 15 modules, MINI allows for the detection of major depressive disorder, dysthymia, suicide risk, manic and hypomanic episodes, panic disorder, agoraphobia, social phobia, generalized

anxiety disorder, OCD, post-traumatic stress disorder, alcohol abuse or dependence, psychotic disorders, anorexia nervosa, and bulimia nervosa.

3. Select a measure for discriminant validity. Select at least one instrument that measures a construct not directly related to the one assessed by the scale that will be validated. As described earlier[34], for validation of the Y-BOCS-II, use two self-report instruments: the Beck Depression Inventory (BDI-II)[39,40], a 21 item self-report questionnaire that assesses the severity of depressive symptoms occurring in the last 15 days; and the State-Trait Anxiety Inventory - Form Y (STAI-Y)[41,42], a 40-item self-report screening instrument developed to measure the severity of anxiety symptoms.

   NOTE: Discriminant validity can be assessed using two distinct approaches: (1) testing against unrelated constructs, where non-significant or very low correlations are expected[43], as demonstrated in this methodology with BDI-II and STAI-Y; or (2) testing against theoretically opposite constructs, where significant negative correlations would be expected[29]. While both approaches are methodologically sound, the present study selected the first approach.

4. Select a measure for convergent validity. Use an instrument that measures the same concept of the scale under validation. To follow the previous example[34], use an instrument that allows the assessment of OCD symptoms, such as the revised Obsessive-Compulsive Inventory (OCI-R)[44].

   NOTE: The OCI-R is an 18 item measure comprising six subscales that cover the full range of OCD symptoms in different settings.

5. Ensure all selected secondary measures are reliable and valid.

   NOTE: It is essential that the instruments selected to assess the psychometric properties of the primary measure are themselves well-established instruments with previously demonstrated reliability and validity[23,25].

3. Translation and cultural adaptation of the primary instrument

1. Perform the back-translation technique[19] to ensure linguistic and semantic equivalence between the original and new versions of the scale.
   1. Perform the forward translation.
      1. Recruit translators. Choose at least two independent translators. Ensure they are bilingual experts in the relevant clinical area (e.g., psychiatrists or psychologists), are native speakers of the language spoken in the country where the study will be conducted, and have expertise in the clinical construct being measured.
      2. Instruct each expert to perform an independent forward translation of the scale from the original version to the new language.
      3. Synthesize and create a consensus forward translation. Instruct the two translators to compare their versions and collaborate to reach a single consensus translation. If disagreements arise that cannot be resolved between the pair, engage a third independent bilingual expert to act as a tiebreaker and facilitate a final consensus.
   2. Perform the backward translation.

1. Recruit new translators. Choose at least two new bilingual translators who are native speakers of the source language of the measure (e.g., English). Ensure they were not involved in the forward translation process to guarantee an unbiased back-translation.
2. Instruct each translator to independently translate the consensus forward translation (from step 3.1.2.1) back into the original language of the scale.
3. Obtain a consensus back-translation by having the two independent back-translators compare and reconcile their versions.

3. Submit the back-translation for author review. Send the consensus back-translation (from step 3.1.2.3) to the authors of the original version of the instrument. Request that they review it for conceptual and semantic equivalence against the original version and provide comments on any discrepancies.
4. Reconcile and finalize the adapted instrument. Ask the initial translation team to compare the consensus back-translation against their consensus forward translation (from step 3.1.1.3). Based on this comparison, and after incorporating the comments from the authors of the original version of the instrument, make final adjustments to the forward translation to produce the pre-pilot version of the adapted instrument.

2. Pilot test the adapted instrument
   1. Administer the pre-pilot version of the instrument to a small sample of at least 10 participants, representative of the target population (e.g., patients with the diagnosis of interest).
   2. Conduct semistructured cognitive debriefing interviews with each participant immediately after administration. Use a standardized guide to gather qualitative feedback on the instrument's clarity, comprehensibility, cultural appropriateness, duration, cognitive effort, and any difficulties encountered during completion.
   3. Systematically review all feedback from both participants and interviewers. Make sure that the research team discusses these inputs to identify any recurring issues or actionable suggestions regarding the instrument's length, the clarity of its items, and the practicality of its format.
   4. Use the inputs from this review to make final evidence-based adjustments, thus creating the final version of the adapted instrument.

4. Selection and recruitment of participants

1. Define the different groups for the recruitment of participants. For a comprehensive validation, recruit participants for three different groups according to the following inclusion criteria:
   1. Group A (Primary Diagnosis Group): recruit participants with the psychiatric diagnosis of interest (e.g., patients with OCD)
   2. Group B (Clinical Control Group): recruit participants with other psychiatric diagnoses relevant to differential diagnosis (e.g., mood or anxiety disorders)
   3. Group C (Healthy Control Group): Recruit healthy volunteers with no current psychiatric disorders.

2. Define inclusion and exclusion criteria.
    1. Define general inclusion criteria. Specify the criteria that all participants across all groups must meet. Ensure that all participants are native speakers of the language spoken in the country where the study will be conducted and do not have conditions or characteristics that compromise the results of the study.
    2. Define general exclusion criteria. List the criteria that will lead to the exclusion of any participant, regardless of their group. To follow the example of Castro-Rodrigues et al.[34] for the validation of a psychiatric instrument such as the Y-BOCS-II, exclude individuals with any of the following: active medical, or specifically, neurological illnesses, such as a clinically significant structural lesion of the central nervous system; acute neuropsychiatric episode that requires hospitalization; history or clinical evidence of chronic psychosis, dementia, developmental disorders associated with low intelligence quotient, or any other form of cognitive impairment; current substance or alcohol abuse or dependence; and illiteracy or inability to understand the study's instructions.
    3. Define group-specific inclusion and exclusion criteria.
        1. For Group A (Primary Diagnosis Group), define the primary inclusion criterion as a confirmed diagnosis of the disorder of interest (e.g., OCD), as determined by a gold-standard diagnostic interview. Typically, no additional exclusion criteria are needed for this group beyond the general ones defined in the previous step (4.2.2).
        2. For Group B (Clinical Controls), define the primary inclusion criterion as a confirmed diagnosis of a relevant psychiatric disorder other than the one of primary interest (e.g., a mood or anxiety disorder). The primary exclusion criterion for this group is a diagnosis of the disorder of interest (e.g., OCD).
        3. For Group C (Healthy Controls), define the primary inclusion criterion as the absence of any current or past psychiatric diagnosis and confirm this via a screening interview. Consequently, the primary exclusion criterion is any evidence of a current or past diagnosed psychiatric disorder.
3. Determine the required sample size.
    1. Conduct an *a priori* power analysis to determine the optimal sample size. Use software, such as G*Power, to calculate the required sample size based on the planned statistical tests (e.g., correlations, *t*-tests), the desired power (typically ≥ 0.80), the alpha level (e.g., 0.05), and the expected effect size based on previous literature[45].

        NOTE: This is the most rigorous method to ensure the study has enough statistical power to detect expected effects.
    2. Ensure the sample size is adequate for factor analysis. While there is no single rule that works for all scenarios, use established guidelines to inform the decision. One common approach is the participant-to-item ratio, with a rule-of-thumb of at least 10 participants per scale item often being

recommended[43]. However, ensure that the final sample size is as large as resources permit, as larger samples lead to lower measurement errors and more stable factor solutions[46].

4. Define the recruitment settings.
    1. For Groups A (Primary Diagnosis) and B (Clinical Controls), recruit participants from an adequate clinical setting for the diagnosis of interest, for example, an outpatient psychiatry clinic where patients eligible for the study are routinely assessed.
    2. For Group C (Healthy Controls), recruit participants through advertisements in public locations likely to reach the same populations and communities that patients belong to.
5. Implement recruitment procedures for each group. For Groups A and B, instruct clinicians collaborating with the study to identify and recruit patients diagnosed with the disorders of interest and willing to participate in the study. Alternatively, randomly identify patients with the diagnoses of interest among patient databases with coding of diagnoses for subsequent recruitment in-person or via telephone.
6. Screen and schedule potential participants. Contact potential participants via telephone, and if they maintain the intention to participate in the study, define a participant ID Code and schedule the first appointment.

5. Preparation and application of the test battery

1. Obtain informed consent from the participant.
    1. Instruct the rater to first assess the participant's capacity to consent, particularly for those with moderate to severe psychiatric symptoms. To do so, evaluate the participant's ability to understand, appreciate, and reason with the study information.
    NOTE: Standardized instruments to evaluate decision-making capacity are available, such as the MacArthur Competence Assessment Tool (MacCAT), a well-established tool for this purpose in populations with psychiatric conditions[47,48].
    2. If a participant is deemed unable to fully understand the information provided, ensure the consent form is signed by their legally authorized representative, if this is stipulated in the protocol and approved by the ethics committee.
2. Standardize the assessment environment. Always conduct the assessment sessions individually in a quiet and private room to minimize distractions and ensure confidentiality. Ensure that each session lasts approximately 60-120 min, depending on the participant's clinical complexity.
3. Administer the initial assessment battery. Following informed consent, instruct the rater (rater A) to administer a clinical questionnaire to assess inclusion and exclusion criteria and to collect other information of interest. If eligibility is confirmed, administer the psychometric instruments in the following order: screening instrument (e.g., MINI), diagnosis instrument (e.g., SCID-IV), other instruments (e.g., STAI, BDI, COI/OCI-R).
4. Handle participant exclusion during assessment. If exclusion criteria are identified at any moment, exclude the participant, thank them for their time, and do not collect additional data.
5. Administer the primary instrument with a blinded rater to evaluate criterion validity. To prevent criterion contamination, instruct a different, blinded rater (rater

B) to administer the primary instrument. Conduct this assessment in the same session or in a second assessment session no more than 1 week after the first. Ensure the second rater is kept blind to results from the first session, in particular, the participant's diagnostic status, by implementing the following procedures:

1. Assign all scheduling and data handling tasks to a member of the research team who is not conducting assessments.
2. Instruct the two raters not to communicate about participants.
3. Provide rater B with only the participant's ID code, ensuring no access to data from the first session.

6. Assess inter-rater reliability. For a subsample or all participants, instruct two different raters to administer the primary instrument in separate sessions, ideally on the same day and not exceeding a 48 h interval, with the order of raters counterbalanced across participants[30,43]. Compare the scores obtained separately by the two raters. A high level of agreement between these scores indicates good inter-rater reliability.
7. Assess test-retest reliability. To evaluate temporal stability, re-administer the primary instrument (e.g., Y-BOCS-II) after an adequate interval, typically 4 weeks, to a subsample or all participants.

6. Statistical analysis

1. Prepare the data for analysis. Use a statistical software package (see the **Table of Materials**) to perform the analysis of psychometric properties.
2. Calculate descriptive statistics. For all sociodemographic, clinical, and psychometric data, calculate descriptive statistics, reporting means and standard deviations for continuous variables and frequencies for categorical variables.
3. Compare group characteristics.
    1. Perform independent samples *t*-tests to compare continuous variables (e.g., age, education, score of the scale under study, and scores of the other psychometric measures), across the different participant groups.
    2. Perform a *Chi-square* ($\chi 2$) test for comparisons of categorical variables, such as gender.
    3. Set the significance level a priori at $p < 0.05$ for all comparisons.
4. Assess the instrument's reliability
    1. Calculate Cronbach's α and McDonald's Ω for the scale and any subscales to evaluate internal consistency. As a hypothesis, define acceptable internal consistency as a Cronbach's α or McDonald's Ω value ≥ 0.70, in line with established guidelines[15].
    2. Calculate the Intraclass Correlation Coefficient (ICC) using data from the test-retest assessments to evaluate temporal stability using Pearson's correlation. Define good test-retest reliability as an ICC value ≥ 0.75, which is considered a strong level of agreement[15].
    3. Calculate Inter-Rater Reliability. Using the scores on the primary instrument collected from the two independent raters (Step 5.5), calculate the Intraclass Correlation Coefficient (ICC) to evaluate inter-rater reliability. An ICC value ≥ 0.75 is

considered evidence of good agreement between raters.

5. Evaluate the instrument's validity.

    1. Assess dimensionality using factor analysis.

        NOTE: The choice of method depends on the existing evidence for the instrument's structure.

        1. If the instrument's factor structure is not yet well-established, perform an Exploratory Factor Analysis by first assessing the suitability of the data for factor analysis using measures like the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. Define acceptable sampling adequacy *a priori* (e.g., KMO > 0.60) and require a significant Bartlett's test ($p < 0.05$)[46]. For items with ordinal response scales (e.g., Likert-type scales), ensure the analysis is conducted on a polychoric correlation matrix using a method such as principal axis factoring with oblique rotation to explore the underlying dimensionality of the items. To determine the number of factors to retain, use multiple criteria, such as the Kaiser criterion (eigenvalues > 1) and examination of the scree plot[46].

        2. If adapting an instrument that already has an established factor structure, perform a Confirmatory Factor Analysis (CFA) to formally test whether the original, established structure is maintained in the newly adapted version. For ordinal data, use an appropriate estimation method, such as Diagonally Weighted Least Squares (DWLS)[49]. As a hypothesis, define acceptable model fit a priori based on established criteria, such as a Comparative Fit Index (CFI) ≥ 0.95, a Tucker-Lewis Index (TLI) ≥ 0.95, and a Root Mean Square Error of Approximation (RMSEA) ≤ 0.06. Evaluate the model fit against these criteria[50].

            NOTE: Performing CFA is particularly useful for examining if the hypothesized structure of the original instrument fits well in the new cultural context.

    2. Assess construct validity.

        1. Calculate Pearson's correlation coefficients to examine the scores of the primary instruments and the scores of the measures selected for convergent and discriminant validity.

        2. To establish convergent validity, hypothesize a significant and strong positive correlation with measures assessing a similar construct.

            NOTE: For example, the PY-BOCS-II[34] total score was expected to correlate strongly with the COI total score.

        3. To establish discriminant validity, hypothesize significantly weaker correlations with measures assessing different constructs than those found for convergent validity. To follow the example of the PY-BOCS-II validation described previously[34], test against unrelated constructs, where non-significant or very low correlations are expected by comparing scores with measures of depression (BDI-II) and anxiety (STAI).

            NOTE: This can also be assessed by testing against theoretically opposite constructs, where

significant negative correlations would be expected.

6. Determine criterion validity for diagnosis.
    1. Hypothesize that the instrument will accurately discriminate between participants with and without the diagnosis of interest.
    2. Use the diagnostic status as defined by the gold-standard instrument selected (e.g., SCID-OCD) as the reference criterion.
    3. Generate a Receiver Operating Characteristic (ROC) curve to study the relationship between scores in the measure of interest and diagnostic status.
    4. Calculate the Area Under the Curve (AUC) to quantify overall diagnostic accuracy. To guide interpretation, define the performance criteria a priori based on established conventions[51], such as: AUC values > 0.90 as excellent, ≥ 0.80 as good, and ≥ 0.70 as fair.
    5. Identify the optimal cut-off value for the score obtained in the measure under study. Identify the score on the instrument that provides the best possible balance between sensitivity and specificity for the intended diagnostic purpose. A common method is to select the value that maximizes the Youden Index (Sensitivity + Specificity - 1)[52].

Representative Results

Despite its gold-standard status and comprehensive structure, the criterion validity of Y-BOCS-II[35] for diagnostic purposes had not been robustly established in the literature at the time the study was conducted. Therefore, this protocol aimed to address this general gap while performing the necessary cultural adaptation for Portugal. In this section, representative data from the validation study of the PY-BOCS-II[34] are presented, with permission from the authors. The results are presented in two parts. First, we summarize the qualitative findings from the pilot testing phase, which informed the final version of the instrument. Second, we present the quantitative results regarding its criterion validity.

The cultural adaptation process included a pilot test with patients, followed by cognitive debriefing interviews to assess the clarity and comprehensibility of the adapted instrument[34]. While most participants found the instructions clear, the qualitative feedback highlighted several key areas for refinement. According to patients, issues included the instrument's length, discomfort with certain examples (reported by one participant), and the difficulty in quantifying the average daily time spent on symptoms due to their episodic nature. The interviewers involved in this process also provided critical feedback, noting that some questions did not flow smoothly, and identified practical formatting issues such as a lack of space for notes and the need for headers to be repeated on each page. This feedback led to adjustments, including minor wording and formatting changes, to produce the final version of the instrument used for the large-scale validation.

For the criterion validity analysis, we recruited a small sample of patients with a diagnosis of either OCD (n = 20) or a mood or anxiety disorder (n = 18), and the PY-BOCS-II was administered by a researcher blinded to the diagnostic status and the results of other psychometric tests, to avoid criterion contamination. Receiver Operating Characteristic (ROC) curves were created to assess criterion validity, using the SCID-OCD as the gold standard for the discrimination between participants with OCD and those with

other diagnoses. **Figure 1** shows the ROC curve for the discrimination between patients with either OCD or another mood and anxiety disorder, assessed in a blinded fashion. An area under the curve (AUC) of 0.93 (95% confidence interval [CI]: 0.84-1.00) was obtained, and further analysis of the ROC curve values demonstrated that a total PY-BOCS-II score of 13 points, when used as a cut-off for diagnosis, correctly identified OCD with a sensitivity of 90% and specificity of 94%. Given the modest sample size and case mix, these accuracy estimates should be replicated in larger cohorts to confirm generalizability.

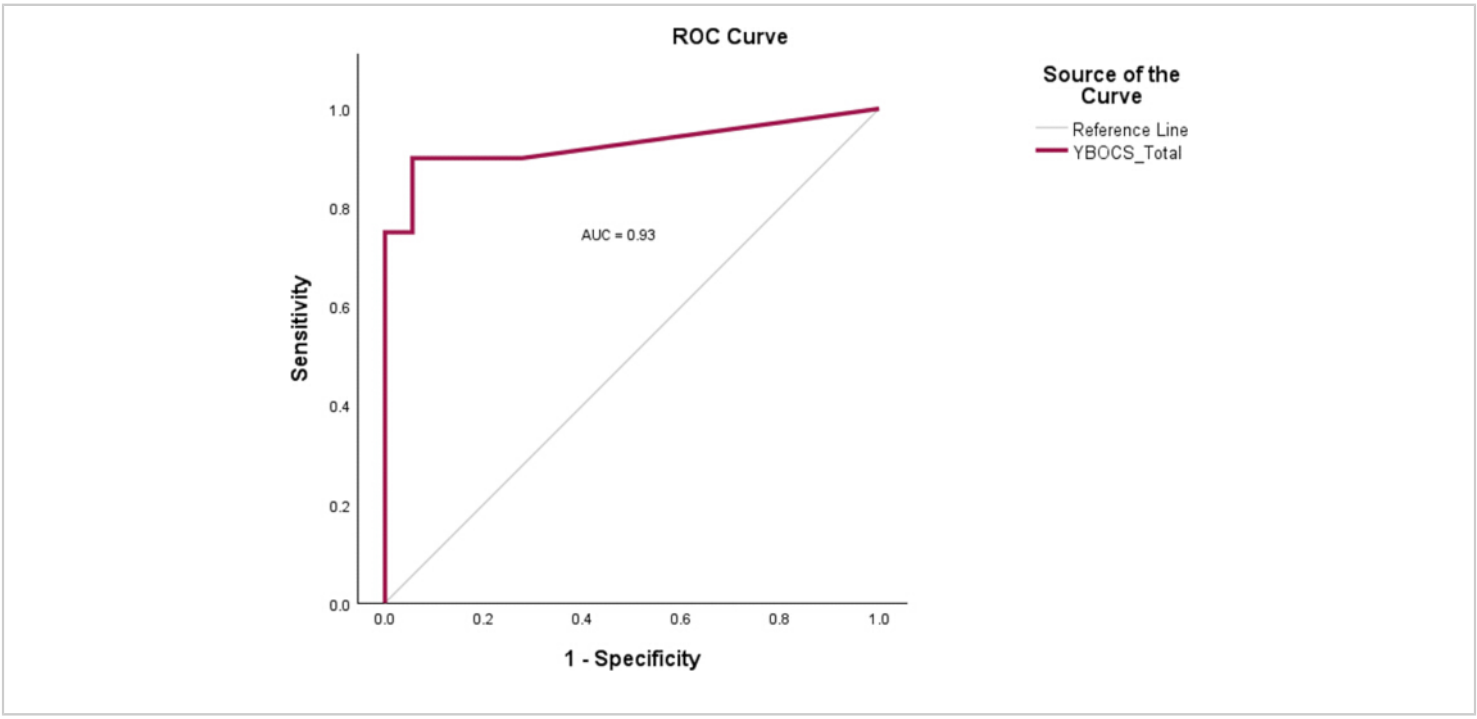

Figure 1: Receiver Operating Characteristic curve for the diagnostic accuracy of the PY-BOCS-II in identifying OCD. This analysis includes data from patients who underwent blinded assessment, comprising a group with OCD (n = 20) and a group with mood and anxiety disorders (n = 18). The plot displays sensitivity (true positive rate) versus 1-specificity (false positive rate) across all possible cut-off scores of the PY-BOCS-II. The SCID-OCD was used as the gold-standard diagnostic tool. Abbreviations: AUC, Area under the curve; OCD, Obsessive-compulsive disorder; PY-BOCS-II, Portuguese Yale-Brown Obsessive-Compulsive Scale-II; ROC, Receiver operating characteristic. This figure was modified from Castro-Rodrigues et al.[34]. Please click here to view a larger version of this figure.

Discussion

Here, we describe a detailed protocol for the cultural adaptation and comprehensive psychometric validation of a psychiatric diagnostic instrument. The protocol begins with the selection of the measure and then details the necessary experimental procedures and statistical analyses. The primary purpose of the protocol is to present a clear and standardized step-by-step procedure to adapt and validate a psychological measure, namely the Y-BOCS-II[34], thereby minimizing confounding factors and undesired variability in

clinical and research use. The methods focus on cultural adaptation and psychometric analysis, including criterion validity, both of which are essential when using an instrument in a new country or context with diagnostic intent[13,17,19].

The Y-BOCS-II[34], a clinician-administered interview for adults, allows for a detailed assessment of OCD. The instrument comprises two main parts: a 67 item Symptom Checklist to identify and classify current and past obsessions, compulsions, and avoidance behaviors, and a 10 item Severity Scale to rate the severity of those symptoms. Despite its gold-standard status and comprehensive structure, its criterion validity for diagnostic purposes had not been robustly established in the literature at the time the study was conducted. Therefore, the protocol aimed to address this broader gap while performing the necessary cultural adaptation for Portugal.

A critical step of the protocol is the rigorous cultural adaptation procedure, performed in line with existing guidelines and evidence-based standards. Equally important is maintaining rater blinding to diagnostic status when applying the psychometric measure, as knowledge of diagnosis can bias outcomes and compromise estimates of diagnostic accuracy[53,54]. This is particularly relevant for structured interviews, as in our example[34]. While compliance with these steps is essential, some features may vary by study (e.g., sample size, item distribution, measurement context, and the attainability of the construct)[55]. In addition, qualitative pilot testing using semi-structured cognitive debriefing with a standardized guide and team review of recurring themes provides actionable evidence for refinement (see Protocol steps 3.2.2-3.2.4). A practical example emerged during the adaptation of item 44: replacing "spouse" with "family member" ensured the instrument captured culturally appropriate reassurance-seeking targets in Portuguese contexts[34].

Beyond translation quality and blinding, comprehensive validation requires principled quantitative assessment of reliability (internal consistency and test-retest stability), construct validity (e.g., factor structure), and criterion validity against an external gold standard[15], following established principles and international guidance, such as COSMIN[59]. For example, for construct validity of the PY-BOCS-II[34], convergent validity was examined against the Coimbra Obsessive-Compulsive Inventory (COI; Inventário Obsessivo de Coimbra)[56], a Portuguese self-report measure with "frequency" and "emotional distress" subscales. While general guidelines for cross-cultural adaptation offer a useful foundation[32], challenges such as overly literal translations and limited stakeholder involvement can compromise the final instrument's validity[57]. In the absence of a single consensus methodology[33], the present protocol provides a transparent, step-by-step framework. Its advantages include mandatory qualitative pilot testing with the target population and blinded-rater assessment to minimize criterion contamination[53], alongside explicit guidance for comprehensive psychometric validation to ensure clinical ethicality[19]. By distinguishing adaptation from validation[33], the protocol is designed to yield a psychometrically sound instrument.

The methodology has been applied successfully in the study by Castro-Rodrigues et al.[34] to assess criterion validity of the PY-BOCS-II clinician-administered interview for diagnosis of OCD. However, the framework is applicable to other formats (e.g., self-report scales and screening questionnaires). For these measures, qualitative pilot testing is paramount to ensure items are unambiguously understood in the

absence of a clinician[58]. Indeed, we have used variations of these methods for other measures and objectives: the Power of Food Scale[60] and Yale Food Addiction Scale[61] (reliability and construct validity), and instruments in oncology settings[62]. Lemos et al.[63] adapted the Perceived Ability to Cope with Trauma Scale, and Almeida et al.[64] adapted the Family Resilience Questionnaire-Short Form (FaRE-SF-P), both incorporating McDonald's Ω alongside Cronbach's α to provide robust internal consistency estimates, especially when tau-equivalence is not met[65]. This consistent methodological approach supports the efficient development of a comprehensive psychometric framework within a given population.

Our criterion-validity protocol has also been effective across clinical contexts. For the Hypomania Checklist-32 (HCL-32)[66], a similar validation protocol was used, with a simplified adaptation process because the measure was already available in Portuguese (Brazilian variant) rather than European Portuguese[67]. The design for that project emphasized screening use in the context of bipolar spectrum disorders, over diagnostic confirmation. More recently, Almeida et al.[68] evaluated the criterion validity of the BDI-II to measure depression severity in patients with cancer, highlighting how somatic symptom overlap can affect diagnostic accuracy.

These applications illustrate the protocol's adaptability across measure types (structured interviews, self-reports), constructs (psychiatric symptoms, appetite-related constructs, mood symptoms, coping, family resilience), and intended uses (diagnosis, screening, severity, psychological resources). With appropriate adjustments to address construct- or population-specific challenges, the protocol can be extended to large-scale screening in primary care and to diverse psychiatric diagnoses. It is also relevant to vulnerable populations where developmental, cognitive, or social factors can influence validity, and to digital health, where mobile-based assessments and digital therapeutics require culturally sensitive validation.

Implementers of this protocol may encounter practical challenges. During translation, if consensus is difficult, involving a senior independent mediator is recommended[69]. Slow recruitment at a single site can be mitigated through multicentre collaboration[70]. For culturally specific content, conceptual adaptation should be prioritized over literal translation, followed by rigorous pilot testing[14,58]. Ethical safeguards are also critical: evaluating capacity to consent, using a legally authorized representative when appropriate, and employing standardized tools (e.g., MacCAT) can support informed participation among individuals with moderate to severe symptoms[47,48].

Limitations include the protocol's resource intensity (time, funding, bilingual experts, trained raters), which may challenge feasibility in low-resource settings. Criterion validation further depends on the availability of a well-established gold standard in the target culture. When such a benchmark is lacking, consensus diagnosis by independent experts is a viable, though demanding, alternative.

In conclusion, this protocol combines multiple empirically supported methods to culturally adapt and validate psychometric instruments for diagnostic use in psychiatry. Its successful applications across diverse instruments show its utility in generating psychometrically sound tools for clinical and research settings across cultural and linguistic contexts.

Disclosures

Albino J. Oliveira-Maia was investigator or national coordinator for Portugal of trials for depression, sponsored by Compass Pathways (EudraCT number 2017-003288-36) and Janssen-Cilag (EudraCT numbers 2019-002992-33, 2022-000439-22, 2022-000430-42); is recipient of a grant from Schuhfried for norming and validation of cognitive tests; has received payment, honoraria, consultancy fees or support for attending meetings and participating in advisory boards from MSD Portugal, Neurolite AG, Janssen-Cilag, the European Monitoring Centre for Drugs and Drug Addiction, Bioprojet Pharma and NaturalX Health Ventures; is Vice President of the Portuguese Society for Psychiatry and Mental Health; is head of the Psychiatry Working Group for the National Board of Medical Examination at the Portuguese Medical Association and Portuguese Ministry of Health; is President of the Ethics Committee for the Portuguese Institute for Addictive Behaviors and Dependence; and is President of the Scientific Council of the Portuguese Obsessive Compulsive Disorder Foundation. None of the aforementioned agencies had a role in the preparation, review, or approval of the manuscript or in the decision to submit the manuscript for publication.

Acknowledgments

This work received funding from the European Union's Horizon research and innovation programme (PsyPal; grant agreement no. 101137378).

References

1. *Mental health: facing the challenges, building solutions: report from the WHO European Ministerial Conference.* World Health Organization, Regional Office for Europe, Copenhagen, Denmark (2005).
2. Prince, M. et al. No health without mental health. *Lancet.* **370** (9590), 859-877 (2007).
3. *ICD-10: International statistical classification of diseases and related health problems.* World Health Organization, Geneva (2011).
4. Maqbul Aljarad, A., Dakhil Al Osaimi, F., Al Huthail, Y. R. Accuracy of psychiatric diagnoses in consultation liaison psychiatry. *J Taibah Univ Med Sci.* **3** (2), 123-128 (2008).
5. National Collaborating Centre for Mental Health (Great Britain), National Institute for Health and Clinical Excellence (Great Britain), British Psychological Society, Royal College of Psychiatrists. *Common mental health disorders: identification and pathways to care.* British Psychological Society; Royal College of Psychiatrists, Leicester; London (2011).
6. *Oxford textbook of correctional psychiatry.* Oxford University Press, Oxford; New York (2014).
7. Beidas, R. S. et al. Free, brief, and validated: standardized instruments for low-resource mental health settings. *Cogn Behav Pract.* **22** (1), 5-19 (2015).
8. Urbina, S. *Essentials of psychological testing.* John Wiley & Sons, Hoboken, NJ (2004).
9. Coulacoglou, C., Saklofske, D. H. *Psychometrics and psychological assessment: principles and applications.* Elsevier/Academic Press, London, United Kingdom (2017).
10. Kimberlin, C. L., Winterstein, A. G. Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm.* **65** (23), 2276-2284 (2008).

11. Roach, K. E. Measurement of health outcomes: reliability, validity and responsiveness. *J Prosthet Orthot.* **18** (6), P8-P12 (2006).
12. Eignor, D. R. The standards for educational and psychological testing. In *APA handbook of testing and assessment in psychology, Vol. 1: test theory and testing and assessment in industrial and organizational psychology.* (Eds) Geisinger, K. F. et al., American Psychological Association, 245-250 (2013).
13. Guillemin, F., Bombardier, C., Beaton, D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol.* **46** (12), 1417-1432 (1993).
14. Beaton, D. E., Bombardier, C., Guillemin, F., Ferraz, M. B. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine.* **25** (24), 3186-3191 (2000).
15. Swan, K. et al. Measuring what matters in healthcare: a practical guide to psychometric principles and instrument development. *Front Psychol.* **14**, 1225850 (2023).
16. Younas, A., Porr, C. A step-by-step approach to developing scales for survey research. *Nurse Res.* **26** (3), 14-19 (2018).
17. ITC guidelines for translating and adapting tests (second edition). *Int J Test.* **18** (2), 101-134 (2018).
18. Hambleton, R. K., Li, S. Criterion-referenced assessment. In *Encyclopedia of statistics in behavioral science.* (Eds) Everitt, B. S., Howell, D. C., John Wiley & Sons, Ltd (2005).
19. Gudmundsson, E. Guidelines for translating and adapting psychological instruments. *Nord Psychol.* **61** (2), 29-45 (2009).
20. Hambleton, R. K., Zenisky, A. L. Translating and adapting tests for cross-cultural assessments. In *Cross-cultural research methods in psychology.* (Eds) Matsumoto, D., van de Vijver, F. J. R., Cambridge University Press, 46-70 (2010).
21. Cook, D. A., Beckman, T. J. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* **119** (2), 166.e7-166.e16 (2006).
22. Domino, G., Domino, M. L. *Psychological testing: an introduction.* Cambridge University Press, Cambridge; New York (2006).
23. Gregory, R. J. *Psychological testing: history, principles and applications.* Pearson Education, Boston (2014).
24. Terwee, C. B. et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* **60** (1), 34-42 (2007).
25. Coaley, K. *An introduction to psychological assessment and psychometrics.* SAGE, Los Angeles (2009).
26. Reith, F. C. M., Van den Brande, R., Synnot, A., Gruen, R., Maas, A. I. R. The reliability of the Glasgow Coma Scale: a systematic review. *Intensive Care Med.* **42** (1), 3-15 (2016).
27. Hubley, A. M., Zhu, S. M., Sasaki, A., Gadermann, A. M. Synthesis of validation practices in two assessment journals: Psychological Assessment and the European Journal of Psychological Assessment. In *Validity and validation in social, behavioral, and health sciences.* (Eds) Zumbo, B. D., Chan, E. K. H., Springer International Publishing, 193-213 (2014).
28. Zumbo, B. D. *Validity and validation in social, behavioral, and health sciences.* Springer, New York (2014).

29. Strauss, M. E., Smith, G. T. Construct validity: advances in theory and methodology. *Annu Rev Clin Psychol.* **5**, 1-25 (2009).

30. Mokkink, L. B. et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. *Qual Life Res.* **18** (3), 313-333 (2009).

31. Mokkink, L. B. et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol.* **63** (7), 737-745 (2010).

32. Sousa, V. D., Rojjanasrirat, W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *J Eval Clin Pract.* **17** (2), 268-274 (2011).

33. Epstein, J., Santo, R. M., Guillemin, F. A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *J Clin Epidemiol.* **68** (4), 435-441 (2015).

34. Castro-Rodrigues, P. et al. Criterion validity of the Yale-Brown Obsessive-Compulsive Scale second edition for diagnosis of obsessive-compulsive disorder in adults. *Front Psychiatry.* **9**, 431 (2018).

35. Goodman, W. K. The Yale-Brown Obsessive Compulsive Scale: II. Validity. *Arch Gen Psychiatry.* **46** (11), 1012-1016 (1989).

36. First, M., Spitzer, R., Gibbon, M., Williams, J. *Structured clinical interview for DSM-IV-TR Axis I disorders, research version, non-patient edition.* DSM-IV, New York (2002).

37. Del-Ben, C. M. et al. Confiabilidade da "Entrevista Clínica Estruturada para o DSM-IV - Versão Clínica" traduzida para o português. *Rev Bras Psiquiatr.* **23** (3), 156-159 (2001).

38. Amorim, P. Mini International Neuropsychiatric Interview (MINI): validação de entrevista breve para diagnóstico de transtornos mentais. *Rev Bras Psiquiatr.* **22** (3), 106-115 (2000).

39. Strunk, K. K., Lane, F. C. The Beck Depression Inventory, second edition (BDI-II): a cross-sample structural analysis. *Meas Eval Couns Dev.* **49** (4), 263-277 (2016).

40. Campos, R. C., Gonçalves, B. The Portuguese version of the Beck Depression Inventory-II (BDI-II): preliminary psychometric data with two nonclinical samples. *Eur J Psychol Assess.* **27** (4), 258-264 (2011).

41. Spielberger, C. D. *Manual for the State-Trait Anxiety Inventory STAI (Form Y).* Consulting Psychologists Press, Palo Alto, CA (1983).

42. Silva, D., Campos, R. Alguns dados normativos do Inventário de Estado-Traço de Ansiedade - Forma Y (STAI-Y), de Spielberger, para a população portuguesa. *Rev Port Psicol.* **33**, 71-89 (1999).

43. Nunnally, J. C., Bernstein, I. H. *Psychometric theory.* McGraw-Hill, New York (1994).

44. Foa, E. B. et al. The obsessive-compulsive inventory: development and validation of a short version. *Psychol Assess.* **14** (4), 485-496 (2002).

45. Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods.* **39** (2), 175-191 (2007).

46. Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. *Multivariate data analysis.* Prentice Hall, Upper Saddle River, NJ (2010).
47. Wang, Y.-Y. et al. The assessment of decision-making competence in patients with depression using the MacArthur competence assessment tools: a systematic review. *Perspect Psychiatr Care.* **54** (2), 206-211 (2018).
48. Wang, S.-B. et al. The MacArthur competence assessment tools for assessing decision-making capacity in schizophrenia: a meta-analysis. *Schizophr Res.* **181**, 104-111 (2017).
49. Li, C.-H. Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. *Behav Res Methods.* **48** (3), 936-949 (2016).
50. Schermelleh-Engel, K., Moosbrugger, H., Müller, H. Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods Psychol Res Online.* **8**, 23-74 (2003).
51. Rice, M. E., Harris, G. T. Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law Hum Behav.* **29** (5), 615-620 (2005).
52. Hughes, G. Youden's index and the weight of evidence revisited. *Methods Inf Med.* **54** (6), 576-577 (2015).
53. Lijmer, J. G. et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* **282** (11), 1061-1066 (1999).
54. Schmidt, R. L., Factor, R. E. Understanding sources of bias in diagnostic accuracy studies. *Arch Pathol Lab Med.* **137** (4), 558-565 (2013).
55. McDonald, R. P. *Test theory: a unified treatment.* Psychology Press, New York (2013).
56. Galhardo, A., Pinto-Gouveia, J. Inventário Obsessivo de Coimbra: avaliação de obsessões e compulsões. *Psychologica.* **48**, 101-124 (2008).
57. Alavi, M., Le Lagadec, D., Cleary, M. Challenges of cross-cultural validation of clinical assessment measures: a practical introduction. *J Adv Nurs.* (2025).
58. Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., Young, S. L. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Front Public Health.* **6**, 149 (2018).
59. Mokkink, L. B. et al. *COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs) user manual.* Department of Epidemiology and Biostatistics, Amsterdam UMC, Amsterdam (2018).
60. Ribeiro, G. et al. Translation, cultural adaptation and validation of the Power of Food Scale for use by adult populations in Portugal. *Acta Med Port.* **28** (5), 575-582 (2015).
61. Torres, S. et al. Psychometric properties of the Portuguese version of the Yale Food Addiction Scale. *Eat Weight Disord Stud Anorex Bulim Obes.* **22** (2), 259-267 (2017).
62. Pettini, G. et al. Predicting effective adaptation to breast cancer to help women BOUNCE back: protocol for a multicenter clinical pilot study. *JMIR Res Protoc.* **11** (10), e34564 (2022).
63. Lemos, R. et al. Cross-cultural adaptation and psychometric evaluation of the Perceived Ability to Cope With Trauma Scale in Portuguese patients with breast cancer. *Front Psychol.* **13**, 800285 (2022).

64. Almeida, S. et al. Cross-cultural adaptation and psychometric evaluation of the Portuguese version of the Family Resilience Questionnaire - short form (FaRE-SF-P) in women with breast cancer. *Front Psychol.* **13**, 1022399 (2022).

65. Dunn, T. J., Baguley, T., Brunsden, V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *Br J Psychol.* **105** (3), 399-412 (2014).

66. Camacho, M. et al. Hypomania symptoms across psychiatric disorders: screening use of the Hypomania Check-List 32 at admission to an outpatient psychiatry clinic. *Front Psychiatry.* **9**, 527 (2018).

67. Soares, O. T., Moreno, D. H., Moura, E. C. de, Angst, J., Moreno, R. A. Reliability and validity of a Brazilian version of the Hypomania Checklist (HCL-32) compared to the Mood Disorder Questionnaire (MDQ). *Rev Bras Psiquiatr.* **32** (4), 416-423 (2010).

68. Almeida, S. et al. Criterion and construct validity of the Beck Depression Inventory (BDI-II) to measure depression in patients with cancer: the contribution of somatic items. *Int J Clin Health Psychol.* **23** (2), 100350 (2023).

69. Eremenco, S. et al. Patient-reported outcome (PRO) Consortium translation process: consensus development of updated best practices. *J Patient Rep Outcomes.* **2** (1), 12 (2018).

70. Gohagan, J. et al. Managing multi-center recruitment in the PLCO cancer screening trial. *Rev Recent Clin Trials.* **10** (3), 187-193 (2015).