Special Collection: Challenges in Translating and Adapting Psychological Measures to Spanish/Portuguese

Original Article

The Bilingual Study Methodology in Translating and Adapting Personality Tests

Equivalence Issues in the Development of the MMPI-2-RF Portuguese Version

Maria João Afonso[1], Rosa Novo[1,2], Cristina Camilo[3,4], and Bárbara Gonzalez[2,5,6]

[1]Faculty of Psychology, University of Lisbon, Portugal
[2]Research Center for Psychological Science, University of Lisbon, Portugal
[3]Iscte - Instituto Universitário de Lisboa, CIS-Iscte, Lisbon, Portugal
[4]CIS-Iscte, Center for Psychological Research and Social Intervention, University of Lisbon, Portugal
[5]School of Psychology and Life Sciences, Lusofona University, Lisbon, Portugal
[6]HEI-Lab: Digital Human–Environment Interaction Labs, Lisbon, Portugal

Abstract: The bilingual samples' studies are listed as a useful tool to confirm the equivalence between linguistically different versions of a test. Yet, such studies are rare in the literature, as they require technical issues to be considered before any conclusion about equivalence can be reached. This paper discusses some of these issues, taking the example of the recent MMPI-2-RF Portuguese adaptation and standardization study. The results of a bilingual study ($N = 53$) using a single-sample design are analyzed, at item, scale, profile, and structural levels, allowing an encouraging general conclusion about the equivalence of the Portuguese MMPI-2-RF to the North American original version, but also pointing out some directions for improvement. The shortcomings of the classical bilingual studies, and the specific limitations due to the obstacles to bilingual samples' recruitment in Portugal, are considered. The limited sample size and some other methodological shortcomings are discussed, considering their implications for future Portuguese MMPI equivalence studies.

Keywords: test translation and adaptation, bilingual study, linguistic/cultural equivalence, personality assessment, Portuguese MMPI-2-RF

Language is the primary means for communicating thoughts and feelings, and for describing behavior. It is long recognized that this feature may interfere with psychological personality assessment, namely, through self-assessment instruments such as inventories or questionnaires. Hence, the task of translating an existing instrument, written in another language, and validated for another population, involves more than simply transposing the words or phrases to the target language. Not only equivalence in verbal content must be considered, but also the cultural context giving sense to the very act of measuring that specific construct, with that specific method, and including those specific contents (item words, phrases, and meanings). The current distinction between test translation and test adaptation, and the consensual option for the latter (International Test Commission [ITC], 2017; Krach et al., 2017), stresses that it is crucial to assure levels of equivalence other than linguistic, like construct and method equivalence (Krach et al., 2017; van de Vijver & Tanzer, 2004). The need for cultural adaptation implies item rephrasing, or even content change, to overcome cultural differences in item reading level and interpretation. This is particularly true when a test written in an Anglo-Saxon language is transposed to a Latin language, as the kind of colloquial wording required to preserve reading level, without interfering in the psychometric and psychological value of each item, often demands substantial verbal change (Krach et al., 2017).

This paper proposes an analysis of the equivalence between the Portuguese adaptation (Novo et al., 2023) and the original Minnesota Multiphasic Personality

Inventory-2-Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008/2011) using a bilingual sample. The inventory's adaptation required significant changes in some items' wording and/or syntax, and the back translation naturally revealed those differences. After their subsequent analysis by the University of Minnesota Press (UMP), and the discussion with the Portuguese research team in view of consensual item improvements, it was necessary to verify the psychometric equivalence between the two versions at different levels (i.e., item, scale, profile, and test levels). The main goal of the adaptation project that included the present study was to make the MMPI-2-RF available to Portuguese psychologists for personality and psychopathology assessment, also allowing its future use in fundamental and applied research, but not specifically in cross-cultural research (Ziegler & Bensch, 2013; see also ITC, 2017).

Theoretical Background

It is almost common sense, in today's psychology, to recognize that behavior and psychological functioning have proximal and distal contexts and cannot be understood, let alone assessed, in a situational, social, or cultural vacuum. At the turn of the 20th century, a contextual/systemic paradigm was already settled down in psychology, paving the way for cross-cultural research and, more specifically, for bias and equivalence research in testing (van de Vijver & Tanzer, 2004), and a list of desirable practices in psychological and educational tests' adaptation (Hambleton, 1994, 2005; Merenda, 2006; van de Vijver & Poortinga, 2005), gave rise to the *ITC Guidelines on Test Adaptation* (2005, 2017).

The certification of item equivalence, not just linguistic content equivalence but rather *test takers' interpretation* equivalence of both versions' items, was considered paramount since the first editions of those guidelines. The use of a bilingual single-sample design is one means for such a confirmation, allowing for the control of the sample's level in what is being measured, while the same sample is exposed to both test versions (Merenda, 2006; Sireci, 2005). Although identified as a research design to test method and item equivalence between the translated and original versions of a test (Hambleton, 1994, 2001, 2005; van de Vijver & Hambleton, 1996; van de Vijver & Tanzer, 2004), some weaknesses were acknowledged from the start: the general population nonrepresentativeness by bilingual samples, due to higher educational level and greater exposure to other cultures, restraining generalization; the difficulty of assuring that participants have, as they should, equal linguistic proficiency and immersion in both cultures; and the practice effect, when tested with the same items in the second version administration (ITC, 2017; Sireci & Berberoğlu, 2000; van de Vijver & Tanzer, 2004).

The confirmation guideline – C-2 (ITC, 2017, p. 18) establishes that "relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations" must be provided, in adapting an existing other language test, and suggests a diversity of specific techniques for independent samples extracted from both populations, fluent in the respective native language and answering to just one version of the test. But the same guideline recognizes that, if "the goal [of test adaptation] is simply to be able to assess test-takers in a different language group on a construct," not to use the test in cross-cultural comparison, "to find evidence of the equivalence of the two forms is not so critical" (ITC, 2017, p. 21), provided that careful validation of the adapted version is done in the new population.

Thus, bilingual samples' studies are still listed today as a strategy for linking (i.e., *equating*) scores across two language versions (confirmation guideline – C-4; relevant for cross-cultural research), but also for evaluating the equivalence of two different language versions of a test (score scales and interpretation guideline – SSI-1; relevant for psychological assessment). Among the diversity of available techniques to perform bilingual studies (Sireci, 2005; Sireci & Berberoğlu, 2000), in this study, a single-sample design was applied: Each participant answers to both the original and translated test versions with about a week to two weeks of interval between administrations.

Beyond the technical design and data analyses issues, it is important to note, as pointed out by Spielberger et al. (2005), that translating and adapting personality or emotional states tests involve specific challenges. Although emotions and personality attributes appear to be universal products of evolution, facilitating equivalence in transcription across languages, subtle obstacles must be considered. First, agreement is not yet achieved in the identification and organization of the diversity of personality and emotional manifestations, even in each culture, let alone among cultures. On the other hand, the same word in different countries, even sharing the same native language (e.g., European and Brazilian Portuguese), may not assume the same meaning due to cultural differences associating different psychological experiences with the same word or phrase. The problem is obviously worse when both native language and culture are diverse. In the personality realm, moreover, the question of item intensity is crucial (e.g., degree of anxiety), both for traits and for states, as each item must preserve the original intensity, with the items in a scale covering various levels

of strength expression. Besides, this intensity or strength of emotional or personality manifestation is also filtered by culture, as an item expressing some degree of emotional experience in one culture may invoke a different feeling in another.

Methods

Participants

The inclusion criteria were (1) Portuguese natives with high proficiency in English, with studies in international English-speaking schools or with a stay of at least 3 years in an English-speaking country; and (2) natives in an English-speaking country, with a good proficiency in Portuguese and living in Portugal for 10 or more years. Participants were recruited either by personal contacts or through the study dissemination.

The initial sample was composed of 63 participants, but some of them were excluded due to missing the second session (seven participants), a number of omissions greater than 10 (two), and VRIN-r (Variable Response Inconsistency-Revised) T > 80 (one). Thus, the final sample was composed by 53 participants aged between 18 and 75 years (M = 33.71, SD = 15.57). The most represented age group was 20–29 years (66%), followed by the age groups of 30–59 years (24%) and 60–75 (10%) years. The participants were mostly women (68%) and were single (64%), married (20%), and divorced (16%). Most of the participants were Portuguese, but 10% of the sample were from the United States, UK, Canada, and South Africa. Proficiency was established on the bases of information conveyed by the participants.

Instruments

Two versions of the Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) were administered, the original North American (NA) version (Ben-Porath & Tellegen, 2008/2011) and the adapted Portuguese (P) version (Novo et al., 2023). The inventory comprises a total of 338 items organized in 51 scales, but only 24 scales were included in this study: 1. seven validity scales – five over-reporting scales: Infrequent Responses (F-r); Infrequent Psychopathology Responses (Fp-r); Infrequent Somatic Responses (Fs); Symptom Validity (FBS); Response Bias Scale (RBS) – and two underreporting: Uncommon Virtues (L-r); Adjustment Validity (K-r); 2. three high-order (H-O) – Emotional/Internalizing Dysfunction (EID); Thought Dysfunction (THD); Behavioral/Externalizing Dysfunction (BXD); 3. nine clinical restructured (RC) – Demoralization (RCd); Somatic Complaints (RC1); Low Positive Emotions (RC2); Cynicism (RC3); Antisocial Behavior (RC4); Ideas of Persecution (RC6); Dysfunctional Negative Emotions (RC7); Aberrant Experiences (RC8); Hypomanic Activation (RC9); and 4. five personality psychopathology (PSY-5) – Aggressiveness-Revised (AGGR-r); Psychoticism-Revised (PSYC-r); Disconstraint-Revised (DISC-r); Negative Emotionality/Neuroticism-Revised (NEGE-r); Introversion/Low Positive Emotionality-Revised (INTR-r). VRIN-r (Variable Response Inconsistency-Revised) and TRIN-r (True Response Inconsistency-Revised) were used to evaluate protocol validity, as they are sensitive to response attitudes specific to each administration.

Each item uses a dichotomic answering format, "true" or "false," and some are inverted (reversed scoring). The administration materials, test booklets, and answer sheets of the Portuguese version are similar to the original version's materials. The answer sheet is prepared for optical reading to allow for the use of an automatic scoring system.

Procedures

A single-group research design was used (Sireci, 2005) involving the administration of the two different language versions to a single group of test takers, most of them within one to two weeks of interval (M = 8.08, SD = 3.07). Although a counterbalanced administration procedure was intended, and generally applied, the final sample included 72% participants who answered to the Portuguese version first. Nine of the 10 excluded participants responded, in the first session, to the English version, suggesting that this version may have been difficult to answer by the Portuguese participants who left the study.

The presentation of the study and the written informed consent always preceded the first administration session. All the administrations were presential, took place in small groups, and followed the administration instructions, also translated, and adapted over the translation process.

Data Analyses

This equivalence study addressed four levels of analysis: item, scale, profile, and whole test (structural) levels. At the item level, true answers (item endorsement) percentages, expressing item *difficulty*, and corrected item-total correlations, expressing item discrimination, were compared between the two versions. Three approaches were used to compare item endorsements between the two

versions: Following the international MMPI-2 adaptation studies (Butcher, 1996), the first procedure compared the two versions' global percentages of true answers, in the whole sample, for each item. A second procedure compared each participant's item endorsement in the two versions, in each item, and analyzed the distributions of −1 (P version replied false and NA version true), +1 (the opposite), and 0 (same endorsement), counting the changes found at individual level (the −1 and +1 differences together); in the third procedure, the McNemar test was used to compare proportions between paired samples, allowing to identify the items where the distribution of the proportions of true/false answers was significantly different between the versions ($p < .05$). A one-parameter latent trait approach was also applied to the substantive scales (Rasch analysis with Winsteps 5.6.4.0 version) to allow for the study of the whole versions' functioning at item level (dimensionality and fitting), and for item and person *logit* measures comparison between versions. To appreciate fitting to the Rasch model, *infit* and *outfit* indices are presented, and the outfit > 2.00 criterium is used to identify item and person misfit.

At the scale level, raw scores were used since differences between the two versions' T scores, obtained with the respective national norms, could be due at least in part to differences between normative data, not to change in test takers' answers. Descriptive statistics and internal consistency indicators were obtained for each scale raw score, for both versions, and intraclass correlation coefficients (two-way random model, absolute agreement) between the two versions were determined, for each of the 24 scales included in the study, alongside with statistical (t test) analyses of mean differences.

At profile level analysis, the substantive scales' profiles, namely, the H-O, RC, and PSY-5, were compared, assessing the agreement between each participant's T scores profile in the two versions of the inventory (T scores, required for profile comparison, were obtained with the respective national norms). First, for each protocol, in the P and NA versions, a profile was identified separately for each group of scales (i.e., H-O, RC, and PSY-5), ranking the respective scales by T scores. Then, the agreement on this ranking, in each profile pair (P version and NA version), was considered according to the following criteria:

- H-O scales: Total Agreement = the three scales follow the same order; Partial Agreement = two of the three scales follow the same order.
- RC scales: Total Agreement = the three highest scales are the same in the two profiles; Partial Agreement = only two of the three highest scales are the same in the two profiles.
- PSY-5 scales: Total Agreement = the two highest scales are the same in the two profiles; Partial Agreement = only one of two highest scales is the same in the two profiles.

Finally, two confirmatory factor analyses (CFA) were conducted for the linguistic versions. The fit indices for the Portuguese and North American versions were compared. Based on the factorial structure of the NA version, the nine restructured clinical scales were organized into three factors: emotional dysfunction, thought dysfunction, and behavioral dysfunction. Model fit was assessed using the root-mean-square error of approximation (RMSEA) and the comparative fit index (CFI; RMSEA values below .08, .05, or .00, and CFI values above .90, .95, or 1.00 demonstrate reasonable, close, and exact fit, respectively). The Tucker–Lewis index (TLI), an adequate index for small samples, was also used (TLI values should be above .90 or .95 for a reasonable or good fit). Comparison of linguistic versions was assessed with $\Delta\chi^2$, ΔCFI, ΔTLI, and ΔRMSEA.

Results

Item Level Equivalence

Table 1 reports item level results within the classical test theory approach. For each of the 24 scales studied, a summary of the two versions' results is presented, and item comparison is based on the range of item results (lowest to highest item observed percentages), and the median for those percentages. In the McNemar test, the number of items where a statistical difference was detected, at least at .05 level, is reported. Finally, the table includes the list of potentially nonequivalent items, identified as having different endorsements between versions by at least two criteria in each scale.

First, item endorsement in the two versions can be appreciated and compared for the 24 scales studied (comparing the minimum–maximum limits of the percentages' range, and the median of the coefficients obtained along the items of each scale). Item *difficulty* coefficients alongside with item discrimination coefficients may be used to compare the psychometric item functioning between the two versions (ITC, 2017). Generally, the two psychometric indices, with a few exceptions, display similar item *difficulty* and item discrimination results for the corresponding scales of the two inventory versions (Table 1). Some of the scales contain items that display very low discrimination coefficients ($r < .30$), as they are rarely endorsed in the general population, such as items describing extreme psychotic

Table 1. Summary of item level analyses by scale: validity, higher-order, restructured clinical, and five personality psychopathology scales

	Portuguese version				North American version				Versions comparison						
	% True		Discrim. coeff.		% True		Discrim. coeff.		Differences in % True[a]		Answering change[b] (% of different answers)			McNemar test	
Scales (Items)	Items range	*Mdn*	Items range	*Mdn*	Items range	*Mdn*	Items range	*Mdn*	Items range	No. of items with dif.> 25%	Items range	*Mdn*	No. of items with > 25% change	No. of items with sig. < .05	MMPI-2-RF items[c]
F-r (32)	0–94	8.0	04–.72	.33	0–96	7.0	.04–.69	.34	−29–11	1	0–32	3.8	1	1	**174**
Fp-r (21)	0–98	8.2	−.13–.55	.28	0–98	6.0	.00–.60	.27	−17–34	1	0–34	1.9	2	2	**79**
Fs (16)	0–92	8.5	−.01–.52	.31	0–92	8.0	−.15–.62	.38	−7–7	0	0–17	5.7	0	0	—
FBS-r (30)	0–87	26.0	− 18–.61	.33	0–87	25.5	.02–.63	.35	−19–34	1	0–34	14.2	2	5	**79**
RBS (28)	0–87	19.0	.02–.52	.29	0–87	19.0	.01–.50	.30	−14–34	1	0–34	10.4	2	1	**79**
L-r (14)	6–94	69.0	−.22–.37	.18	8–96	75.0	.03–.55	.19	−15–14	0	6–24	13.2	0	2	—
K-r (14)	34–85	48.0	.09–.67	.42	32–77	53.0	.17–.60	.36	−7–15	0	11–37	20.8	3	1	—
EID (41)	6–100	45.0	.15–.73	.46	6–94	45.0	.04–.75	.48	−32–17	1	0–36	15.1	4	2	**25**, 140, 167
THD (26)	0–87	5.0	.04–.58	.38	0–91	6.0	.06–.66	.38	−12–6	0	0–24	2.9	0	0	—
BXD (23)	0–92	21.0	−.01–.47	.27	0–91	19.0	.16–.53	.32	−14–17	0	0–21	7.5	0	2	—
RCd (24)	13–92	29.0	.03–.73	.50	8–75	28.0	.23–.75	.50	−13–41	1	6–47	17.0	1	1	**130**
RC1 (27)	2–94	62.0	.02–.61	.35	2–96	70.0	−.00–.62	.31	−29–12	1	0–34	9.4	4	3	**174**
RC2 (17)	26–100	64.0	.19–.51	.35	28–94	75.0	.04–.50	.28	−32–14	1	0–36	15.1	3	4	**25**, 140, 195
RC3 (15)	23–85	46.0	.04–.57	.41	28–83	47.0	.21–.76	.42	−13–15	0	2–36	17.0	2	1	—
RC4 (22)	0–92	19.0	−.01–.47	.21	0–91	18.0	.16–.52	.32	−19–17	0	0–37	4.9	2	2	—
RC6 (17)	0–87	6.0	.30–.67	.49	0–91	6.0	.27–.70	.52	−4–15	0	0–26	1.9	1	0	—
RC7 (24)	4–85	32.0	.14–.68	.41	2–79	31.0	.05–.72	.38	−7–12	0	2–27	13.2	2	1	—
RC8 (18)	0–75	14.0	.04–.57	.32	0–87	12.5	.06–.61	.36	−12–9	0	0–19	8.5	0	0	—
RC9 (28)	0–83	42.5	.09–.44	.28	2–77	39.0	.17–.59	.31	−10–32	2	2–40	15.1	4	3	**143, 181**
AGGR-r (18)	0–87	45.5	.18–.51	.31	2–92	47.5	.18–.56	.32	−9–14	0	2–27	13.2	1	1	—
PSYC-r (26)	0–75	4.0	.04–.58	.39	0–87	6.0	.06–.66	.34	−12–6	0	0–24	3.8	0	0	—
DISC-r (20)	4–92	25.5	−.01–.47	.33	4–91	23.0	.16–.53	.36	−14–17	0	0–21	6.6	0	2	—
NEGE-r (20)	4–85	46.0	.14–.60	.34	2–79	40.5	.05–.63	.35	−6–17	0	2–28	18.9	4	2	167
INTR-r (20)	23–100	66.0	.15–.60	.34	17–92	64.0	.04–.59	.36	−19–30	1	6–38	15.1	4	2	140, 153, **181**, 195

Note. Item endorsement/difficulty (percentage true) and item discrimination (corrected item-total correlation). Differences between the Portuguese and North American versions (N = 53). [a]Difference between the global percentages of participants answering true in each item, without consideration of change by participant. Negative differences refer to items' endorsements (true responses) higher in the North American version. [b]Percentage of participants who changed item endorsements between the two versions. [c]List of items identified by at least two of the three criteria (Answering change > 25% and McNemar test with sig. < .05). In bold, items identified by all three criteria.

symptoms, such as delusions or hallucinations. These items display the same low rate of endorsement in the North American population (Butcher et al., 2001), and their modest discrimination power is not just a consequence of translation but rather of low variance in general population samples. Table 1 also displays the results for the statistical comparison of items' endorsements between the two versions.

Pertaining to the first criterium to identify items with equivalence problems, 14 out of the 24 scales showed no item with a significant difference (> 25%) in endorsements between the two versions. The remaining 10 scales contained just one or two items with differences in percentages of endorsement higher than 25% (28.3%–46.7%), but some of those items were included in several scales. In the final count, only seven items (out of 273) displayed nonequivalent endorsements, for this first criterium, although a few presented high percentage differences (e.g., Item 130, 46.7%).

With the second criterium, the number of items identified with more than 25% of participants changing the endorsement between versions was higher, 22 items, including the seven items already detected by the previous procedure. Finally, with the third criterium, 21 items displayed significant differences at least at $p < .05$ level, 10 of which were also identified by the second criterium, and seven were identified by all three criteria. In sum, the 10 items identified by at least two of the three criteria for the nonequivalence analyses, and listed in Table 1 (column MMPI-2-RF items), were the primary object of revision, but they represent only 3.6% of all items in the 24 scales included in this study.

In a Rasch analysis (Table 2 and Table 3), although several results suggest a tendency toward scale multidimensionality (e.g., the median eigenvalue for the first contrast in the residuals is 2.9 for both versions, ranging between 2.1 and 5.0 in P version, and between 2.2 and 5.2 in NA version), the percentage of total variance explained by the model measures always exceeds the 20% criterium (Reckase, 1979; median of 31.3 and 35.4, for the P and NA versions, respectively, ranging between 22.7% and 52.4% in the P version, and between 21.2% and 43.2% in the NA version), supporting the latent trait analysis procedure with these data. The dimensionality analysis at the item level (Table 2) allows comparing the two versions first contrasts identified by a factor analysis of the residuals.

The percentage of variance explained by the model is generally similar in the P and NA versions, although some differences emerge, in the number of contrasts that may be interpreted as a dimension, with eigenvalues above 2.00 (Linacre, 2023). Although these results must be interpreted cautiously, due to specificities of this kind of analysis that explores contrasts in the data, not latent constructs (Linacre, 2023), it is worth noting the general presence of one to two more expressive contrasts in both versions' scales, and the whole resemblance at scale level variance structure.

Table 3 reports the summary of the Rasch item analysis of the substantive scales H-O, RC, and PSY-5. As could be expected in a general population sample, means for person measures are generally low, well below the mean of item difficulty (by convention, the point 0 of the logit scale). Some extreme low measures led to the automatic exclusion of items from the model, justifying item positive means, something probably less common in clinical samples. In both versions, maximum infit values are all below 2.0, most of them even below 1.5 (all means near 1.00 in both versions), while maximum outfit indices allow identifying some misfit in items and persons. However, both item and person misfit percentages are well below 10% for both versions (medians of misfit percentages were 3.7% for items and 3.8% for persons in both versions), suggesting a general and similar fit of the two versions to the Rasch model.

The high to very high intraclass correlations between the logit measures of the two versions' items, .83–.98 (*Mdn* .95), and persons, .80–.95 (*Mdn* .88), alongside with other similarities between item functioning of the two versions, support their general equivalence. The list of items displaying misfit, in any of the versions, is not similar to the list of items identified as differently replied in the two versions, in endorsement analysis, meaning that the differently endorsed items are not necessarily misfit items in Rasch analyses of both versions. As an exception, Item 130 stands out: The one with the largest change in endorsement across the versions in classical item analysis is also the one with the more severe misfit in the P version (outfit = 9.90, well above the NA version outfit = 2.14). This pointed out the absolute need for that specific item translation revision.

Scale and Profile Level Equivalence

The intraclass correlation coefficients reported in Table 4 were systematically high or very high (all above .80, about half above .90) in all sets of scales. The paired samples comparisons showed that, in the majority of the 24 scales (about 75%), the mean scale score was the same in the two versions, as few significant statistical differences were found (only two in the 17 substantive scales considered). The item internal consistency coefficients (Cronbach's α) were acceptable to high, especially in the substantive scales, and at about the same level in both versions, in almost all the scales, showing a similarity in items' psychometric functioning within the scales of both versions. This is important to note as some of the lowest α coefficients in the P version are coincident with the corresponding coefficients

Table 2. Rasch contrast analysis (exploratory factor analysis of the residuals) of the substantive scales: high-order, restructured clinical, and five personality psychopathology scales

	Portuguese version							North American version						
	Model	1st contrast		2nd contrast		3rd contrast		Model	1st contrast		2nd contrast		3rd contrast	
Scales (Items)	% Total variance	Eigenvalue	% Total variance	Eigenvalue	% Total variance	Eigenvalue	% Total variance	% Total variance	Eigenvalue	% Total variance	Eigenvalue	% Total variance	Eigenvalue	% Total variance
EID (41)	34.8	4.95	8.1	3.39	5.5	3.01	4.9	36.5	5.16	8.0	3.44	5.3	2.89	4.5
THD (26)	23.4	3.01	11.5	2.60	9.9			21.5	3.66	14.4	2.73	10.7	2.03	8.0
BXD (23)	30.3	2.97	10.4	2.19	7.7			35.4	2.86	8.8	2.65	8.2	2.02	6.2
RCd (24)	43.0	3.30	7.8	2.42	5.8	2.09	5.0	38.9	3.44	8.8	2.38	6.1		
RC1 (27)	29.8	3.08	8.0	2.74	7.1	2.43	6.3	22.4	3.28	9.4	2.73	7.9	2.28	6.6
RC2 (17)	31.3	2.39	10.3	2.02	8.7			31.9	2.47	9.9	2.15	8.6		
RC3 (15)	39.6	2.10	8.5					39.7	2.47	9.9				
RC4 (22)	22.7	3.12	12.1					31.9	2.63	8.9	2.35	8.0		
RC6 (17)	52.4	3.11	11.4					42.9	3.83	16.8				
RC7 (24)	39.1	2.49	6.3	2.42	6.1	2.21	5.6	37.8	2.43	6.3	2.32	6.0		
RC8 (18)	23.3	2.19	10.5					22.1	2.02	9.8				
RC9 (28)	28.1	2.91	7.8	2.66	7.1	2.55	6.8	30.3	3.45	8.6	2.63	6.5	2.43	6.0
AGGR-r (18)	35.8	2.91	11.0	2.30	8.7	2.01	7.6	43.2	3.00	9.5				
PSYC-r (26)	24.1	2.86	10.9	2.59	9.8	2.34	8.9	21.2	3.65	13.7	2.30	8.6		
DISC-r (20)	28.8	3.07	10.9					31.0	2.74	9.5	2.21	7.6	2.01	6.9
NEGE-r (20)	37.0	2.88	9.1	2.24	7.1			35.9	2.33	7.5				
INTR-r (20)	31.4	2.93	10.6	2.16	7.8			37.7	2.65	8.2	2.28	7.1		

Note. Variance explained by the model and by each of the first three contrasts in the Portuguese and North American versions (N = 53). Only contrasts with Eigenvalue > 2.00 (representing at least a two items dimension) were retained.

Table 3. Summary of Rasch analysis of the substantive scales items: high-order, restructured clinical, and five personality psychopathology scales

	Portuguese version								North American version								Correlations	
Scales (Items)	Item measures[a] *M* (*SD*)	Person measures[b] *M* (*SD*)	Item infit Min–Max	Item infit *M* (*SD*)	Item outfit Min–Max	Item outfit *M* (*SD*)	Item misfit[c] *f*/%	Person misfit[c] *f*/%	Item measures[a] *M* (*SD*)	Person measures[b] *M* (*SD*)	Item infit Min–Max	Item infit *M* (*SD*)	Item outfit Min–Max	Item outfit *M* (*SD*)	Item misfit[c] *f*/%	Person misfit[c] *f*/%	Item meas. *r*[d]	Person meas. *r*[d]
EID (41)	.13 (1.38)	−.84 (1.40)	.64–1.53	.99 (.20)	.42–2.29	.95 (.39)	1/2.44	3/5.66	.00 (1.17)	−1.06 (1.48)	.60–1.45	.99 (.23)	.37–2.75	1.01 (.58)	3/7.32	1/1.89	.91	.90
THD (26)	.68 (1.67)	−3.09 (1.37)	.74–1.22	1.00 (.61)	.13–2.31	.98 (.61)	2/7.69	4/7.55	.71 (1.60)	−3.04 (1.44)	.64–1.41	.99 (.79)	.29–1.63	.90 (.44)	0/0	2/3.77	.95	.88
BXD (23)	.54 (1.84)	−1.66 (1.25)	.73–1.25	1.00 (.14)	.36–2.59	.97 (.49)	1/4.35	4/7.55	.36 (1.87)	−1.81 (1.35)	.79–1.30	.98 (.15)	.15–1.98	.84 (.43)	0/0	2/3.77	.98	.91
RCd (24)	.00 (1.40)	−1.31 (1.88)	.54–1.41	1.00 (.23)	.30–9.90	1.24(1.88)	1/4.17	4/7.55	.00 (1.01)	−1.51 (1.84)	.49–1.53	.99 (.25)	.26–2.14	.95 (.52)	1/4.17	2/3.77	.83	.91
RC1 (27)	.00 (1.40)	−2.41 (1.43)	.50–1.31	.98 (.19)	.05–2.03	.86 (.40)	1/3.70	3/5.66	.00 (1.21)	−2.42 (1.19)	.69–1.33	.98 (.19)	.31–2.31	.89 (.47)	1/3.70	2/3.77	.92	.80
RC2 (17)	.30 (1.65)	−.79 (1.33)	.78–1.26	1.01 (.15)	.64–1.27	.94 (.20)	0/0	1/1.89	.00 (1.31)	−1.18 (1.25)	.77–1.17	1.00 (.11)	.62–1.31	.95 (.19)	0/0	4/7.55	.88	.83
RC3 (15)	.00 (1.36)	.05 (1.73)	.49–1.75	.98 (.29)	.41–1.92	.95 (.44)	0/0	1/1.89	.00 (1.10)	−.05 (1.85)	.59–1.58	.98 (.26)	.42–2.96	1.09 (.68)	1/6.67	4/7.55	.95	.93
RC4 (22)	.32 (1.50)	−.22 (1.30)	.78–1.27	1.00 (.12)	.31–2.05	.93 (.36)	1/4.55	2/3.77	.36 (1.66)	−2.2 (1.44)	.81–1.20	.98 (.11)	.58–2.13	.96 (.39)	1/4.55	4/7.55	.96	.80
RC6 (17)	.97 (2.56)	−3.76 (1.97)	.41–1.87	1.01 (.47)	.05–2.44	.83 (.63)	1/5.88	2/3.77	.94 (2.20)	−3.38 (1.62)	.29–1.40	.96 (.35)	.08–2.11	.82 (.69)	1/5.88	2/3.77	.98	.88
RC7 (24)	.00 (1.48)	−1.31 (1.65)	.69–1.23	.99 (.89)	.15–1.68	.94 (.39)	0/0	2/3.77	.00 (1.55)	−1.51 (1.71)	.60–1.32	.99 (.20)	.34–2.02	.99 (.44)	1/4.17	2/3.77	.97	.85
RC8 (18)	.36 (1.55)	−2.79 (1.43)	.74–1.15	1.02 (.11)	.46–2.39	1.06 (.46)	1/5.60	4/7.55	.38 (1.43)	−2.72 (1.38)	.84–1.37	1.01 (.17)	.43–2.29	1.08 (.44)	1/5.6	4/7.55	.91	.84
RC9 (28)	.19 (1.49)	−.38 (.89)	.80–1.31	.99 (.10)	.44–1.31	.97 (.18)	0/0	0/0	.00 (1.31)	−.72 (1.02)	.70–1.46	.99 (.15)	.14–1.60	.93 (.94)	0/0	0/0	.95	.89
AGGR-r (18)	.33 (2.01)	.16 (1.21)	.81–1.24	1.00 (.14)	.47–1.62	.96 (.30)	0/0	2/3.77	.00 (1.91)	−18 (1.24)	.78–1.16	1.00 (.10)	.23–1.79	.97 (.31)	0/0	4/7.55	.98	.91
PSYC-r (26)	.69 (1.70)	−3.21 (1.48)	.75–1.30	1.01 (.16)	.15–2.89	.92 (.63)	1/3.85	2/3.77	.59 (1.54)	−2.92 (1.33)	.65–1.25	1.00 (.17)	.29–1.81	.93 (.43)	0/0	4/7.55	.95	.88
DISC-r (20)	.00 (1.13)	−1.72 (1.39)	.71–1.31	.99 (.15)	.46–2.78	1.02 (.55)	1/5.00	3/5.66	.00 (1.22)	−1.67 (1.38)	.76–1.49	.98 (.18)	.41–1.85	.91 (.39)	0/0	3/5.66	.96	.95
NEGE-r (20)	.00 (1.45)	−.68 (1.43)	.70–1.33	1.00 (.16)	.60–1.59	.97 (.27)	0/0	2/3.77	.00 (1.53)	−.88 (1.35)	.72–1.23	1.01 (.17)	.43–1.30	.93 (.29)	0/0	2/3.77	.97	.86
INTR-r (20)	.26 (1.57)	−.53 (1.17)	.75–1.21	1.00 (.15)	.64–1.43	1.00 (.26)	0/0	1/1.89	.00 (1.35)	−.77 (1.53)	.71–1.34	.99 (.17)	.57–2.48	1.11 (.46)	1/5.00	5/9.43	.87	.86

Note. [a]All items (including extreme score items). Mean values different from .00 indicate that at least one item was excluded from the model due to extreme score. [b]All persons (including extreme score persons). [c]Outfit MNSQ (mean-square fit statistics) > 2.0. Items. Portuguese version (8 items): 57i (EID), 273 (THD and RC8), 122 (THD), 237i (BXD, RC4 and DISC-r), 130 (RCd), 52i (RC1), 212i (RC6), and 92 (PSYC-r). North American version (11 items): 25i, 102i, 140i (EID), 130 (RCd), 52i (RC1), 238 ((RC3)), 5 (RC4), 310 (RC6), 132 (RC7), 330 (RC8), and 246i (INTR-r). Intraclass correlations (two-way random model, absolute agreement) between item and persons logit measures for the Portuguese and North American versions (N = 53). [d]$p < .001$ for all correlations.

Table 4. Scale level analyses – validity, higher-order, restructured clinical, and five personality psychopathology scales: raw scores' descriptive statistics for the Portuguese and North American versions

	Portuguese version			North American version			*t* test		Intraclass correlations between versions	
Scales (Items)	*M*	*SD*	α	*M*	*SD*	α	*t* (*df* = 52)	*p*	*r*	95% CI
F-r (32)	3.60	3.68	.82	2.64	2.98	.77	3.61	< .001	.89	[.77–.94]
Fp-r (21)	2.38	2.00	.63	1.96	1.93	.64	2.37	.022	.87	[.77–.93]
Fs (16)	1.28	1.77	.67	1.25	1.63	.61	0.21	.832	.84	[.72–.91]
FBS-r (30)	8.77	3.71	.67	8.08	3.30	.57	2.56	.013	.91	[.83–.95]
RBS (28)	6.85	3.40	.69	6.09	3.00	.61	2.82	.007	.89	[.79–.94]
L-r (14)	4.06	1.96	.45	3.68	2.14	.57	1.68	.098	.81	[.67–.88]
K-r (14)	6.34	3.17	.72	6.55	3.16	.72	−0.73	.467	.88	[.80–.93]
EID (41)	14.89	8.81	.92	14.02	9.01	.92	1.71	.094	.95	[.92–.97]
THD (26)	2.19	2.44	.73	2.13	2.66	.78	0.23	.817	.87	[.77–.92]
BXD (23)	4.94	3.21	.74	5.11	3.37	.76	−0.74	.460	.93	[.88–.96]
RCd (24)	7.57	5.88	.91	6.92	6.13	.92	1.65	.106	.94	[.89–.97]
RC1 (27)	4.79	4.03	.82	4.13	3.45	.76	1.91	.062	.87	[.77–.92]
RC2 (17)	6.08	3.32	.74	5.40	2.81	.64	2.50	.016	.87	[.77–.93]
RC3 (15)	7.49	3.74	.83	7.23	4.14	.86	1.06	.296	.94	[.90–.97]
RC4 (22)	3.51	2.58	.66	3.79	3.21	.78	−1.01	.318	.86	[.76–.92]
RC6 (17)	1.58	2.27	.84	1.40	2.33	.86	1.14	.261	.93	[.87–.96]
RC7 (24)	7.72	5.13	.87	7.26	5.00	.86	1.21	.232	.92	[.86–.95]
RC8 (18)	2.15	2.26	.70	1.96	2.35	.76	.86	.396	.86	[.76–.92]
RC9 (28)	11.55	4.32	.73	10.64	4.93	.80	2.67	.010	.92	[.84–.95]
AGGR-r (18)	9.19	3.04	.68	8.89	3.14	.73	1.44	.156	.93	[.89–.96]
PSYC-r (26)	2.17	2.61	.77	2.34	2.59	.74	−0.76	.450	.89	[.81–.94]
DISC-r (20)	4.87	3.44	.77	5.00	3.49	.77	−0.69	.496	.96	[.93–.98]
NEGE-r (20)	8.15	4.12	.80	7.58	4.02	.79	1.64	.106	.89	[.82–.94]
INTR-r (20)	7.74	3.83	.77	7.79	4.21	.81	−.20	.846	.93	[.87–.96]

Note. Paired samples differences (*t*) and intraclass correlation coefficients (*r*). (*N* = 53).

in the NA version results. Then, some scales lower internal consistencies of the P version seem not to be due to translation flaws. Hence, the high intraclass correlation coefficients, together with the few significant differences between the two versions scores means, exhibit high coherence between the scale level results obtained by the same persons in the two inventory versions.

At the profile level, the results of the P and NA versions, paired by participant, showed adequate agreement in the different groups of scales (Table 5). The agreement at the profile level, between the two versions of the inventory, suggests that the clinical interpretation would be very similar across the two versions. The fact that the RC scales comprise a high number of scales (nine) contributes to lower degree of total agreement than the one found in the H-O and PSY-5 scales, with only three and five scales, respectively. However, in 95% of the cases, profile agreement was present in at least two of the three highest RC scales.

Furthermore, in neither group of scales, a complete disagreement between the profiles of the two versions was observed, i.e., there is always agreement at least in one of the two or three highest scales in each group.

Test Level Structural Comparison

Finally, Table 6 displays standardized factor loadings for the P and NA versions, and model fit indices for both linguistic versions, obtained via a CFA.

Significant factor loadings of approximately the same size showed similar patterns across the two versions. Model fits, measured via RMSEA, CFI, and TLI, were good for both versions. Chi-squared comparison of the models showed that the Portuguese and North American versions are not significantly different from each other. In addition, the ΔCFI, ΔTLI, and ΔRMSEA

Table 5. Profile level analyses: higher-order, restructured clinical, and five personality psychopathology scales

Groups of substantive scales	Total agreement (*f*) %	Partial agreement (*f*) %	Without agreement (*f*) %
Higher-order (H-O) (3 scales)	(34) 81	(10) 19	(0) 0
Restructured clinical (RC) (9 scales)	(31) 59	(19) 36; (3) 5[a]	(0) 0
Personality psychopathology scales (PSY-5) (5 scales)	(46) 87	(87) 13	(0) 0

Note. Percentage of agreement between the Portuguese and North American versions (paired samples, $N = 53$). [a]In three cases (5%), the partial agreement was observed in only one of the three most elevated RC scales.

Table 6. Standardized factor loadings and model fit indices for the Portuguese and North American versions of the nine restructured clinical (RC) scales in a three-factor model ($N = 53$)

Factors RC scales	Portuguese version	North American version
Emotional dysfunction		
RCd	.82***	.72***
RC1	.64***	.54***
RC2	.61***	.45**
RC7	.96***	.93***
Thought dysfunction		
RC3	.73***	.71***
RC6	.80***	.68***
RC8	.58***	.59***
Behavioral dysfunction		
RC4	.89***	.75***
RC9	.72***	.81***
CFI	.99	.90
TLI	.99	.81
RMSEA	.03	.08
χ^2 (*df*)	21.30 (20)	23.09 (19)
ΔCFI	.09	
ΔTLI	.18	
ΔRMSEA	−.05	
$\Delta\chi^2$ (*df*)	1.79 (1), $p = .157$	

Note. ***$p < .001$. **$p < .01$.

results also indicate similar adjustments for the two versions.

Discussion

This study was part of a larger research project, the adaptation and standardization of the MMPI-2-RF to Portugal. The present analysis of equivalence between the Portuguese and North American versions of the MMPI-2-RF started by a back translation of an initial translated version, and its subsequent comparison with the original inventory. This previous phase of the project culminated in a consensual P version, approved by the University of Minnesota Press (UMP). Next, the present bilingual study tried to contribute to establish the equivalence between the P and the NA versions' measures (scales, profiles, and internal structure) and between current psychometric indicators (item analysis, dimensionality, and reliability coefficients).

With the first criterium used to analyze item level equivalence (Butcher, 1996), seven items (7/273) were identified as deserving attention due to the percentage (> 25%) of change in endorsement between the two versions. The second criterium identified a list of 22 items with more than a 25% change across versions, which included the previous seven items, and in the third procedure (McNemar test), 21 items were identified as significantly different between versions. Finally, 10 items (less than 4%) were pointed out by at least two of the three criteria and were considered the ones deserving a critical revision attention.

Latent trait approach to item analyses provided a means for comparing paired samples dimensionality, as well as item and person fit to the Rasch model and logit measures. The confirmation of a central dimension, in all the scales of both versions, with the model explaining a significant proportion of the total variance, and the emergence of a first contrast (sometimes also a second), similar between versions in eigenvalue and explained variance, supports in some way the equivalence between versions. As expected, in a nonclinical sample, person measures means were low (below item difficulty mean), as item measures displayed higher level, in the logit scale, in both versions. Infit and outfit indices tended to achieve the same approximate levels in both versions while, above all, intraclass correlations, used to control random effects affecting differently the two moments of administration, displayed very high results, above .90 in several scales (82% for item measures and 35% of person measures). Some item misfit probably justifies the not perfect correlations between item and person logit measures.

Internal consistency α coefficients were generally high for both versions, in the main substantive scales, and located at the same approximate level in both versions. Item analysis coefficients also displayed similar level of item functioning in both versions, and even when the lower

limit of the coefficients' range is at a low level in the P version, the same holds for the NA version, due to low variance items containing extreme psychopathology contents, rarely endorsed in the general population.

At scale level analyses, most of the scales presented no significant differences in the mean values between the two versions – only six out of 24 scales with mean scores statistically different and just one scale at $p < .001$. Taking the main substantive scales, only two significant differences were found for RC2–Low Positive Emotions ($p < .016$) and RC9–Hypomanic Activation ($p < .010$), in which item means were lower in the NA version. This means that this sample scores in the substantive scales were generally the same in both versions. The scale scores intraclass correlations were high or even very high, further supporting the general equivalence of the versions at scale level: All scale correlations were significant at $p < .001$, and all the main substantive scales presented correlations above .80, and about half, even above .90, including the high-order EID–Emotional/Internalizing Dysfunction scale (.95) and the RCd–Restructured Clinical Demoralization scale (.94).

The high level of profile equivalence, between the two versions of the inventory, is satisfactory, with high total agreement for the H-O (81%) and PSY-5 (87%) scale groups, and mean values for the RC scales group (59%). In the RC scales, the partial agreement in the profiles (in two of the three highest scales) reaches 95%. In fact, the configurational interpretation of the results, as emphasized by the MMPI literature, is diagnostically richer and more useful than the interpretation without regard for the relationships among the scales. The level of agreement achieved indicates that both versions allow the generation of very similar diagnostic hypotheses or psychological inferences about the examinee.

Considering measurement factorial structure, a similar pattern was observed across the P and NA versions, and in both versions, CFA resulted in a good fit. This suggests that the latent variables are related to the restructured clinical scales in the same way for both versions.

A careful analysis of item content and comparison between the two versions was the final step in this equivalence study, in view of necessary and appropriate item reviews (Krach et al., 2017; Merenda, 2006). As an example, one nonequivalent item, between versions, by all three criteria, that also showed serious misfit in the Rasch analysis of the P version, Item 130, was compared with the original regarding its content. While the NA version used a known metaphorical (idiomatic) expression for describing a tendency to ruminate or to stay bothered, in the Portuguese translation, the phrase, containing no metaphor, was somehow ambiguous due to lack of context. The P version item could be interpreted as a sign of thoughtfulness or intelligence, while the original one, on the contrary, was more specific also presenting low social desirability, as ruminating is not considered a healthy coping strategy even by lay persons. In this item, 44% of the participants who gave an answer chose true in the P version and false in the NA version. In trying to overcome this lack of equivalence, this item was then modified to also include a current Portuguese metaphor with a sense much closer to ruminating, and similar social desirability. Indeed, this item is a good example of the type of challenge faced in translating personality tests (Spielberger et al., 2005): Translating idiomatic expressions is especially hard, and adaptation is needed to translate the feeling connotation of the original idiom, and its exact intensity, rather than the literal meaning of the words. Ideally, similar idiomatic expressions should be used, with the same emotional or psychological connotation and identical strength, while trying to avoid interfering in the psychometric item functioning.

Limitations and Conclusion

The bilingual study, although using the quite unusual single-sample design (Sireci, 2005), was useful for an approach to the equivalence between the original NA and the P adapted versions of the MMPI-2-RF. The general conclusion is in favor of a general linguistic, psychological, and psychometric equivalence, at multiple levels, and the few items identified for revision were reconsidered and improved in the final MMPI-2-RF. Yet, the shortcomings of the single-group bilingual studies must be kept in mind, as well as the use of bilingual samples assuming adequate and similar fluency in both languages, and equal immersion in both cultures. Future studies must objectively evaluate proficiency and collect more information about life experience in each cultural context. Another weakness of this study was sample dimension, considering the high number of items composing some scales, preventing from structural analyses at the item level. Although the main purpose of all analyses in this single-sample equivalence study means that the same sample shortcomings affected both versions under comparison in a similar way, future validation studies should test measurement invariance with larger samples. Some demographic asymmetries, a consequence of the difficulty of recruiting bilingual participants, as well as the failure to carry out the counterbalanced administration procedure also affected methodological options, drawing attention to improvements that must be kept in mind in future Portuguese MMPI bilingual studies.

References

Ben-Porath, Y. S., & Tellegen, A. (2008/2011). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2 Restructured Form): Manual for administration, scoring, and interpretation*. University of Minnesota Press.

Butcher, J. N. (Ed.), (1996). *International adaptations of the MMPI-2. Research and clinical applications*. University of Minnesota Press.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2 (Minnesota Multiphasic Personality Inventory-2): MMPI-2 manual for administration, scoring, and interpretation* (rev. ed.). University of Minnesota Press.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*(3), 229–244.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, *17*(3), 164–172. https://doi.org/10.1027/1015-5759.17.3.164

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Lawrence S. Erlbaum Publishers.

International Test Commission [ITC]. (2005). *International guidelines on test adaptation*. https://www.intestcom.org/page/14

International Test Commission [ITC]. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). https://www.intestcom.org/page/14

Krach, S. K., McCreery, M. P., & Guerard, J. (2017). Cultural-linguistic test adaptations: Guidelines for selection, alteration, use, and review. *School Psychology International*, *38*(I), 3–21. https://doi.org/10.1177/0143034316684672

Linacre, J. M. (2023). *A user's guide to Winsteps/Ministeps Rasch-model computer programs (Program manual 5.6.3)*. https://www.winsteps.com/a/Winsteps-Manual.pdf

Merenda, P. F. (2006). An overview of adapting educational and psychological assessment instruments: Past and present. *Psychological Reports*, *99*(2), 307–314. https://doi.org/10.2466/pr0.99.2.307-314

Novo, R. F., Afonso, M. J., & Gonzalez, B. (2023). *MMPI-2-RF – Inventário Multifásico de Personalidade de Minnesota-2 Forma Reestruturada* [Adaptation of the MMPI-2-RF – Minnesota Multiphasic Personality Inventory-2 – Restructured Form]. Hogrefe.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207–230. https://doi.org/10.2307/1164671

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117–138). Lawrence S. Erlbaum Publishers.

Sireci, S. G., & Berberoğlu. (2000). Using bilingual respondents to evaluate translated /adapted items. *Applied Measurement in Education*, *13*(3), 229–248. https://doi.org/10.1207/S15324818AME1303_1

Spielberger, C. D., Moscoso, M. S., & Brunner, T. M. (2005). Cross-cultural assessment of emotional states and personality traits. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 343–367). Lawrence S. Erlbaum Publishers.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*(2), 89–99. https://doi.org/10.1027/1016-9040.1.2.89

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135. https://doi.org/10.1016/j.erap.2003.12.004

van de Vijver, F., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Lawrence S. Erlbaum Publishers.

Ziegler, M., & Bensch, D. (2013). Lost in translation: Thoughts regarding the translation of existing psychological measures into other languages. *European Journal of Psychological Assessment*, *29*(2), 81–83. https://doi.org/10.1027/1015-5759/a000167

History

Received July 4, 2023
Revision received February 10, 2024
Accepted February 16, 2024
Published online April 23, 2024

Section: Methodological Topics in Assessment

Open Science

The information needed to reproduce all the reported methodology is available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Maria João Afonso
https://orcid.org/0000-0001-5555-5677
Rosa Novo
https://orcid.org/0000-0002-4670-6987
Cristina Camilo
https://orcid.org/0000-0002-0767-445X
Bárbara Gonzalez
https://orcid.org/0000-0001-5142-256X

Bárbara Gonzalez

School of Psychology and Life Sciences
Lusófona University
Campo Grande, 376
1749-024 Lisbon
Portugal
barbara.gonzalez@ulusofona.pt