Diagnostic accuracy of a Brazilian depression self-report measure (EBADEP): Original and short versions

Makilim Nunes Baptista[1], *Lucas de Francisco Carvalho*
Universidade São Francisco, Campus Swift, Campinas-SP, Brasil

ABSTRACT

The present study aimed to investigate the diagnostic accuracy of a scale for the assessment of depressive symptoms, the EBADEP-A, and its short version. Also, we propose the application of the concept of equivalence to compare the discriminative capacity between the two scales, i.e., to determine whether a new procedure (i.e., EBADEP-A short version) is not worse than the procedure in use (i.e., the EBADEP-A). Participants were 80 individuals, 40 with a diagnostic of major depressive disorder based on SCID-I, and 40 healthy controls. Age mean was about 41, and most participants were female (85%). We calculated sensitivity, specificity, positive and negative likelihood ratio, positive and negative predictive power, and negative predictive power (PP−), based on the ROC curve for the EBADEP-A full and short versions. Both versions of EBADEP-A proved useful for the diagnostic of major depression. We concluded that the short version of EBADEP-A is closely equivalent to its extended version. Based on the results we found, we conclude that using diagnostic accuracy combined to the concept of equivalence can not only assist professionals with a focus on psychometric parameters, but also have an impact on the application of the tools in clinical settings.

Keywords: psychometric properties; diagnostic; cut-off.

RESUMO – Acurácia diagnóstica de uma medida de autorrelato para avaliação da depressão (EBADEP): versões original e breve

O presente estudo teve como objetivo investigar a acurácia diagnóstica de um instrumento para avaliação de sintomas depressivos, a EBADEP-A, e sua versão breve (EBADEP-A versão breve). Além disso, nós propomos a aplicação do conceito de equivalência para comparar a capacidade discriminativa das duas escalas, isto é, para determinar se o novo procedimento (i.e., EBADEP-A versão breve) não é inferior ao procedimento já utilizado (i.e., EBADEP-A). Os participantes foram 80 sujeitos, sendo 40 com diagnóstico de transtorno depressivo maior com base na SCID-I e 40 sem esse diagnóstico. A média de idade foi aproximadamente 41 e a maioria dos participantes foram mulheres (85%). Nós calculamos a sensibilidade, especificidade, o *likelihood ratio* positivo e negativo, poder preditivo positivo e negativo, baseados na curva ROC para as versões completa e breve da EBADEP-A. Ambas as versões da EBADEP-A apresentaram resultados favoráveis para o uso na prática diagnóstica. Nós concluímos que a versão breve da EBADEP-A foi equivalente à sua versão completa. Com base nos resultados encontrados, parece que a proposta apresentada, qual seja, usar a acurácia diagnóstica combinada ao conceito de equivalência, pode não somente auxiliar os profissionais com foco nos parâmetros psicométricos, mas também tendo impacto nas aplicações clínicas da ferramenta.

Palavras-chave: propriedades psicométricas; diagnóstico; ponto de corte.

RESUMEN – Precisión diagnóstica de una medida de autoinforme para evaluación de la depresión brasileña (EBADEP): versiones originales y cortas

El presente estudio tuvo como objetivo investigar la precisión diagnóstica de un instrumento para la evaluación de síntomas depresivos, el EBADEP-A, y su versión corta (EBADEP-A versión corta). Además, se propuso la aplicación del concepto de equivalencia para comparar la capacidad discriminatoria entre las dos escalas, es decir, para determinar si el nuevo procedimiento (i.e. EBADEP-A versión corta) no es inferior al procedimiento en uso (i.e. EBADEP-A). Al total participaron 80 individuos, de los cuales 40 poseían el diagnóstico de trastorno depresivo mayor basado en SCID-I y 40 sin este diagnóstico. El promedio de edad fue de aproximadamente 41 años y la mayoría de los participantes eran mujeres (85%). Se calculó la sensibilidad, la especificidad, el *likelihood ratio* positivo y negativo, el poder predictivo positivo y negativo, y el poder predictivo negativo (PP-), basados en la curva ROC para las versiones completa y corta de EBADEP-A. Ambas versiones de EBADEP-A mostraron resultados favorables para su uso en la práctica diagnóstica. En conclusión, la versión corta de EBADEP-A fue equivalente a su versión completa. Con base en los resultados obtenidos, parece que la propuesta presentada, al utilizar la precisión diagnóstica combinada al concepto de equivalencia, no sólo puede ayudar a los profesionales con un enfoque en los parámetros psicométricos, sino que también puede tener un impacto en las aplicaciones clínicas de la herramienta.

Palabras clave: propiedades psicométricas; diagnóstico; punto de corte.

Depression is considered a public health concern. Depressive disorders are highly prevalent and increasing in the general population, including physical and psychological symptoms, high comorbidity with

[1] Endereço para correspondência: Universidade São Francisco. Rua Waldemar César da Silveira, 105, Jardim Cura D'Ars (SWIFT), 13045-510, Campinas, SP. *E-mail:* makilim01@gmail.com

other medical conditions, disability, and premature death. This was the conclusion of an epidemiological study, part of the Mental Health Survey conducted by the World Health Organization (WHO), which assessed 18 countries, emphasizing Brazil as the country with the highest prevalence rate of Major Depressive Disorder (Bromet et al. 2011) among the low- and middle-income countries analyzed.

The first step in the treatment of depression is assessment. Although diagnostic criteria are well operationalized in psychiatric manuals (e.g., DSM-5 and ICD-11), alternative depression scales exist that can assist in assessment procedures such as diagnosis, screening, patient follow-up, and measurement of treatment outcomes. However, for these instruments to function effectively in aiding the diagnostic process, it is recommended that their diagnostic accuracy should be investigated (Kroenke, 2001; Mitchell & Coyne, 2007; Pirraglia, Rosen, Hermann, Olchanski, Neumann, & 2004).

Diagnostic accuracy describes the capacity of a test to identify true and false cases of a disease. It includes two main concepts, sensitivity (i.e., the correct identification of true positives), and specificity (i.e., the correct identification of true negatives), as well as other indicators such as positive and negative likelihood ratio, which indicate the probability of presence or absence of the disease when the test is positive or negative, respectively (Parshall, 2013; Zhou, Obuchowski, & McClish, 2002). Individuals correctly classified as possessing the disorder are named true positives, while those identified correctly as not possessing the disorder are called true negatives. Similarly, healthy individuals misclassified as having the disorder are referred to as false positives, and those affected but misclassified as not having the disorder are referred to as false negatives. All this information can be obtained when a diagnostic reference standard (called "gold standard") is available (Derogatis & Lynn, 2000; Iared & Valente, 2009).

Sensitive tests (i.e., those with high sensitivity) are useful when the detection of a disease is the primary purpose of assessment, including the screening of a disease at early stages, when several concurrent possibilities should be discharged. In turn, specific tests (i.e., those with high specificity) are often used as collateral sources to confirm the diagnostic as indicated by other measures. On the one hand, the ideal diagnostic potential occurs when a test is highly sensitive and specific. A cut-off is then established at the point that best favors the tradeoff between high sensitivity and specificity. On the other hand, in the case of screening tests, when one cannot run the risk of identifying people as healthy when they have the disorder, sensitivity should be preferred over specificity, which might increase false positives, but also minimize the occurrence of false negatives (Andreoli, Blay & Mari, 1998; Fletcher, Fletcher & Wagner, 1996; Klein & Costa, 1987). Moreover, screening tests must be less costly and time-consuming when compared to longer tests, even though they might be a bit less specific.

In Brazil, some scales for the assessment of depression symptoms are available that can be used in clinical practice by psychologists (i.e., tests approved by the National Psychological Testing System – SATEPSI). Instances include the Beck Depression Scale (BDI), the second version of the BDI (BDI-II) and the *Escala Baptista de Depressão* (Baptista Depression Scale) – Adult version (EBADEP-A). Many studies support the psychometric properties of all these scales for use in Brazil. Cunha (2001) investigated the accuracy of the BDI, and found 77% of sensitivity and 95% of specificity in a sample of 148 patients with major depressive disorder and 148 with dysthymic disorder with groups paired by age, and using a psychiatric diagnosis as gold standard. Gomes-Oliveira, Gorenstein, Lotufo-Neto, Andrade, and Wang (2012) found, for the BDI-II, 70% of sensitivity and 87% of specificity in a non-clinical sample of 182 adults, including 60 male, and using the SCID-I as gold standard. Baptista (2012) investigated the accuracy of the EBADEP-A, having found 77.5% of sensitivity and 87.5 of specificity in a sample of 1,676 adults in which 527 presented clinically significant depression symptoms according to the BDI. All the cited scales present sound psychometric properties, even though they might differ in their targeted population, item response format and other characteristics.

EBADEP-A is the longest instrument, containing 45 items that cover a higher number of clinical descriptors, while the other two have only 21 items each (Baptista, 2012; Cunha, 2001; Goreinstein, Wang, Argimon, Werlang, &2011). However, the EBADEP-A has the items with the smallest number of words, besides a response format that affords completion in equal or less time when compared to the two versions of the BDI. The use of abbreviated scales, called "short versions," has been usual, since they are less costly and save time, provided that they meet psychometric quality standards, such as sensitivity and specificity values near those of the longer or reference scales (i.e., gold standards). Moreover, according to Bolsoni and Zuardi (2015), screening scales for mental disorders with fewer items may be very useful in primary health care.

A few short version of well-known scales can be found in the literature, such as the Hamilton Depression Rating Scale (HAM-D) and the Center for Epidemiologic Studies Depression (CES-D). Interestingly, the Symptom Checklist-core Depression (SCL-CD), albeit consisting of only six items, achieved a sensitivity of 76% and specificity of 96% in a representative sample of approximately 6,000 participants in Sweden (Hanson et al., 2014). Bech et al. (2009) also tested a six-item short version of the HAM-D, in a sample of 153 Israelis, and compared with the diagnostic criteria of the Structured Clinical Interview for

DSM (SCID), they found 100% sensitivity and 91% specify, i.e., excellent indices for that specific sample. Baron, Davies and Lund (2017) assessed a 10-item version of the CES-D in an African sample of 944 participants, reporting sensitivity of 84.6% and specificity of 84% for the general sample, with a positive predictive value of 53.7% representing that this percentage of people with the cut-off of 11 points or more established in the ROC curve presented a diagnostic of depression by clinical criteria. Björgvinsson, Kertz, Bigda-Peyton, McCoy and Aderka (2014) assessed 755 psychiatric patients who were participants in a hospital intervention program using a 10-item version of CES-D, finding sensitivity of .89 and specificity of .47. This last index is relatively low, so that the authors argue that the scale is probably adequate to screening and the assessment of severity of depressive symptoms, even though it is not recommended for diagnostic use. Despite containing fewer items, short versions can reach fair sensitivity and specificity indices and, in some cases, prove to be as accurate as their full versions. Of course, it should be considered that scales with a smaller number of items (e.g., six) might not cover the entire spectrum of depressive symptomatology, then failing to assess core clinical indicators (Akena et al., 2012).

It is essential that studies with a comparative design be implemented to test the equivalence in diagnostic capacity of short and full versions of the scales for assessing depression. In medicine, comparative studies are commonly applied for checking whether one intervention is better than another, or whether the procedures are clinically and statistically different (Greene, Morland, Durkalski, & Frueh, 2008; Lesaffre, 2008). However, the concepts from comparative studies – i.e., superiority, equivalence, and non-inferiority – do not seem to be used in the context of diagnostic accuracy of mental health instruments, in which comparisons would not be performed between assessment tests. Usually, these concepts are applied for comparison between intervention procedures to determine: 1. unilaterally whether a new procedure is worse than another already in use (non-inferiority); 2. bilaterally whether a new procedure is not inferior to another procedure already established (equivalence); or 3. unilaterally whether a new procedure is better than another procedure already in use (superiority).

The present study deals with the diagnostic accuracy of an instrument for the evaluation of depressive symptoms, the EBADEP-A, and its short version (EBADEP-A short version). We propose the application of the equivalence concept, traditionally applied in the case of comparisons between interventions in medicine, to compare the discriminative capacity between the two scales, i.e. to determine whether a new procedure (i.e., EBADEP-A short version) is not worse than the procedure in use (i.e., the EBADEP-A full version). Although better results for the short version of the scale would be a desirable outcome, this study has no *a priori* hypotheses as this is the first comparison between the scales.

Method

Participants

The study included 80 individuals from the countryside of the state of São Paulo, 40 with a diagnostic of major depressive disorder based on SCID-I and 40 without this diagnostic according to the same interview. The groups were matched by sex and age, and the age ranged from 22 to 69 years, both for the clinical group (M=40.53, SD=12.32) and for the healthy controls (M=40.58, SD=12.31). Both groups presented the same sex distribution (85% female).

Instruments

Escala Baptista de Depressão (Adult version) - EBADEP-A (Baptista, 2012). This instrument was designed and standardized in Brazil, with the aim of tracking symptoms of depression in psychiatric and non-psychiatric samples. The scale was developed based on the Diagnostic and Statistical Manual of Mental Disorders – DSM-IV-TR (APA, 2002), the International Classification of Diseases - ICD-10 (WHO, 1993), the Cognitive model (Beck, Rush, Shaw, & Emery, 1997), and the Behavioral Theory (Ferster, Culbertson, & Boren, 1977). This one-dimension scale consists of 45 items disposed on a semantic-differential format, containing two contrasting statements each. A rating scale is used to measure if the participant agrees more with the first statement, the second or with both at the same magnitude. Studies using Classical Test Theory and Item Response Theory procedures have revealed sound psychometric properties for the scale in the screening of depression, with Cronbach's alpha reliability of 0.94, and 0.92 of real item precision found according to the Rasch model (Baptista, 2012).

For the short version of EBADEP-A, 15 items were selected by the authors of this study to match the major depression descriptors most commonly used in psychiatric manuals (i.e., core symptoms, APA, 2014). The items 1, 2, 4, 5, 10, 18, 23, 24, 28, 29, 31, 38, 40, 43, and 45 were chosen, as they capture humor features, anhedonia, guilt, fatigue, concentration, suicidal ideation, and sleep, among others. Correlations between items ranged from .11 to .60 (M=.29; SD=.09), and internal consistency (α) was .86, suggesting the short version was similarly unidimensional and precise.

Structured interview for DSM-IV - Clinical version (SCID-CV). Adapted to Portuguese by Del-Ben et al. (2001), this interview was built based on an extended research version of the SCID. In the current study, we employed 15 questions that screen the main

DSM-IV criteria for mood disorders. A reliability study was conducted by Del-Ben et al. (2001) in psychiatric inpatients in the countryside of the state of São Paulo. The test-retest methodology was used, with a two-day interval between interviews. Participants included 45 patients, with a mean age of 34.9 years (SD=11.8), most of them women (60%). The between-rater agreement regarding diagnostic (Kappa) was higher than 0.90, with significance at 1%, which led to the conclusion that the scale has good reliability, despite not presenting all the criteria included in the full research version of the interview.

Procedures

The project of this study was submitted and approved by an Institutional Review Board (CAAE: 0422.0.142.000-11). Administrations with patients were individual. Data collection took place in two different places, in a private psychiatric clinic (which also received patients who had medical insurance) and in a public health center. In the clinic, the collection was conducted after the patients' medical appointment, in a room assigned for such application. Only patients with the diagnostic of Major Depressive Disorder were referred by the psychiatrist (who used ICD-10 criteria for diagnostic classification). In the Health Center, the cases of depression were referred by a multi-professional team (physiotherapist, psychiatrist, psychologists and health agents). Depressive patients who received a home visit from the Family Health Program (PSF) team were invited to participate in the study and, in these cases, the administration was performed at the patients' homes. With the depressive patients, in general, the objectives of the research were first explained, and the patient signed an Informed Consent Form (ICF). Then, the SCID-CV interview for mood disorders was applied. After confirmation of Major Depressive Disorder, the characterization form and the EBADEP-A were filled. In total, 67 depressed patients were interviewed, of whom 12 discontinued the administration and 15 did not confirm the diagnosis of major depression (probably because they were under medication, and the main symptoms had already disappeared). The administration in each patient lasted, on average, an hour and a half. All data were collected throughout approximately nine months. Finally, with the data of the depressive patients in hand, we searched, in universities, for the same number of people with an unconfirmed diagnosis for depression, with paired sex, age, and education. The data collection also occurred individually, according to the availability of place and time of each person. After signing the consent form, the first two questions of the SCID-CV (referring to the central symptoms of depression) were asked, and the absence of these symptoms allowed the continuation of the collection with the other instruments.

Data analysis

ROC curve was applied for inspecting sensitivity and specificity of the full and the short versions of the EBADEP-A. We also calculated positive likelihood ratio, negative likelihood ratio, positive (PP+) predictive power, and negative predictive power (PP−), and the 2 × 2 tables (see Parshall, 2013). The PP+ and the PP− were calculated according to the formulas presented by Streiner (2003), for cases in which the samples do not present prevalence rates according to the population (i.e., 7%, APA, 2013). To investigate the equivalence between the EBADEP-A versions, the data obtained in the 2 × 2 tables were compared using kappa and intraclass correlation (ICC) indices, and the r^2 provided by the logistic regression analysis (VI: test scores; DV: dichotomous variable being a patient or not). It is worth noting that the Kullback (1959) proposal was used to compare the coefficients (r^2) obtained.

Results

According to the objectives of this research, the ROC curve was used to investigate the diagnostic capacity of EBADEP-A, besides the comparison between the full and short versions of the scale. Figure 1 shows the Area Under the Curve (AUC) for the two versions of EBADEP-A.

The curves for the full and short versions of EBADEP-A resulted very similar. Thus, both AUCs found were equal to 0.98. Regarding the full version, the optimal cut-off seems to be the score equal to 48, which yields a sensitivity of 97.5% and specificity of 90%. Table 1 displays the distribution of patients and non-patients according to the gold standard (columns) and according to the EBADEP-A full version.

It can be observed a greater occurrence of false positive cases in comparison to false negatives, which is expected given the highest sensitivity relative to specificity. Complementing these results, the positive and negative likelihood ratios (LR+ and LR−, respectively) were also calculated, which were equal to 0.03 (95%CI: 0-0.19) and 9.73 (95%CI: 3.83-24.69), indicating a fair confidence in identifying subjects who do not have the disorder and a moderately good confidence for subjects who have the disorder. Regarding the positive and negative predictive power, the first was equal to 0.98 and the second equal to 0.93, suggesting that a person scoring positive in EBADEP-A has a 98% probability of having a depressive condition, and a person with a negative result in the test has a 93% probability of not having depressive symptoms.

Next, the optimal cut-off for the short version of EBADEP-A was also verified, given the ROC curve, which suggested an optimal score of 19. Table 2 shows the distribution data for the short version of EBADEP-A.

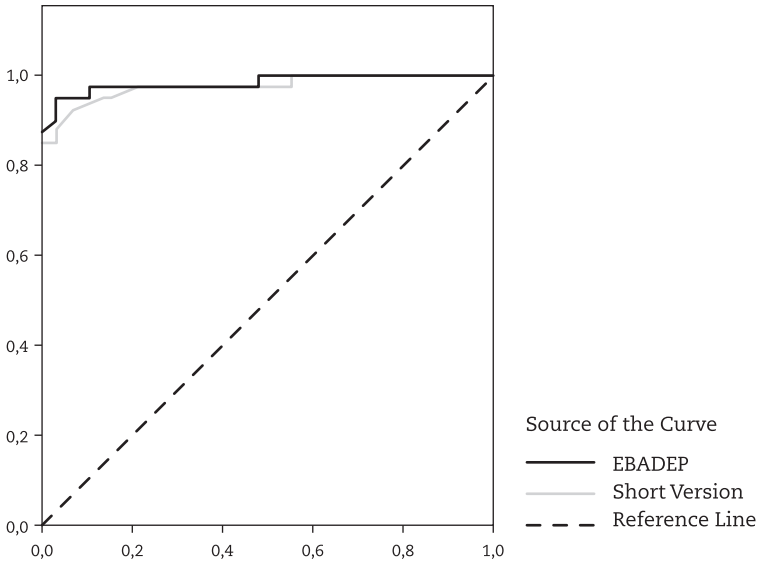

Figure 1. AUC of the EBADEP-A Versions

Note. continuous line=EBADEP-A full version; continuous and dotted line=EBADEP-A short version; dotted line=reference line (AUC=.50)

Table 1
Distribution of Patients and Non-Patients Using the Cut-Off for the Full Version

	Patients and non-patients	
	Depressive	Non-depressive
EBADEP-A +	39	4
EBADEP-A –	1	36

Table 2
Distribution of Patients and Non-Patients Using the Cut-Off for the Short Version

	Patients and non-patients	
	Depressive	Non- depressive
EBADEP-A_{short} +	38	5
EBADEP-A_{short} -	2	35

Table 2 demonstrates that the short version can correctly predict practically all patients with the disorder, and a slightly larger number of people who are not diagnosed as depressed. These numbers are reflected in the sensitivity indexes (95%, 95% CI: 83.08-99.39) and specificity (87.5%, 95% CI: 73.20-95.81), so that the first was slightly higher than the second. Complementary, LR+ was equal to 7.60 (95% CI: 3.34-17.31) and LR- equal to 0.06 (95% CI: 0.01-0.22). Similarly to the full version of EBADEP-A, these indices suggest a generally good confidence in classifying subjects who do not have the disorder and moderately good for subjects who have the disorder. Regarding the positive and the negative predictive power, the first was equal to 0.94 and the second equal to 0.88, suggesting that a person who is positive in EBADEP-A has a 94% probability of having a depressive condition, and a person who is negative in the test has a 88% probability of not having depressive symptoms.

Finally, to compare the results obtained for the ability in detecting true positives and true negatives, that is, the equivalence between the two EBADEP-A versions, Table 3 presents the distribution of people with and without a diagnosis of depression.

Table 3
Equivalence Test by Kappa

		EBADEP-A short version		Total
		Test –	Test +	
EBADEP-A full version	Test –	36	1	37
	Test +	1	42	43
Total		37	43	80

Note. Test-=cut-off for depression was not reached; Test+=cut-off for depression was reached

We observed that 90% of the patients with no diagnostic of depression were correctly allocated by both versions of EBADEP-A. Similarly, most of the patients diagnosed with depression by the gold standard were also correctly allocated by the instruments; however, 5% of the sample with no depression according to the gold standard were allocated as scoring above the cut-off in the EBADEP-A versions. Again, the full version presents a small gain over the short version; however, both instruments present a high level of agreement as indicated by the Kappa index ($k=0.95$, $p<0.001$) and the intraclass correlation (ICC=0.92; $p<0.001$). Complementary, when using logistic regression analysis, we observed a $r^2_{nagelkerke}=0.89$ ($p<0.001$) for the full version of the EBADEP-A and $r^2_{nagelkerke}=0.85$ ($p<0.001$) for the short version, coefficients that are not significantly different ($p=0.38$) based on Kullback's proposal (1959).

Discussion

Despite the expressive number of assessment instruments for mood disorders developed worldwide specifically for depression (Santor, Gregus & Welch, 2006), the number of people diagnosed may be much lower than the real cases diagnosed. For example, in a meta-analysis performed in more than 50,000 cases and 118 diagnostic accuracy studies, in the primary health setting, health professionals were able to correctly diagnose only 47.3% of the cases, which shows low diagnostic accuracy (Mitchell, Varze & Rao, 2009).

Despite the existence of some depression scales with well-established diagnostic accuracy results (Baron et al., 2017; Bech et al., 2009), it is important to emphasize that new scales may be useful for assessing the symptomatology of depression, once these scales might be constituted by symptoms that are not directly reported in psychiatric manuals. The EBADEP-A, in its adult version, assesses the main symptoms described in psychiatric manuals and other sources, addressing the main psychosocial theories of depression, such as hopelessness, self-esteem, and pessimism, among others. In addition, item format follows a semantic-differential design, also evaluating patients' positive characteristics and containing a percentage of items in pre-established dimensions (Calil & Pires, 1998), different from most of the scales used in Brazil (Baptista, 2012).

It is also important to note that, for some disorders, only the detection of symptoms present in the psychiatric manuals may not be adequate for the accurate diagnostic, the reason why it is important to use complementary scales and screening (Soares, Moreno, Moura, Angst, & Moreno, 2010). In this sense, the objective of this study was to compare the diagnostic accuracy of two versions of EBADEP-A, since accuracy is one of the most important psychometric characteristics of a diagnostic test (Andreoli et al., 1998; Fletcher et al., 1996).

In the present study, both versions of EBADEP-A showed very promising results for their use in diagnostic practice when compared to the main results of reference instruments regarding sensitivity and specificity. For example, except for the study conducted by Bech et al. (2009) with the six-item HAM-D, which found very high values for sensitivity and specificity, even outside the standard observed by other measures, these indices were higher for both the EBADEP-A versions when compared to many other published similar scales (e.g., BDI-I and II, CES-D, SCL).

The results showed that both versions reached the cut-off above 0.80 in the AUC, which is desirable (Hajian-Tilaki, 2013; Obuchowski, 2000). Likewise, the occurrence of false positives and false negatives in both versions reached quite acceptable standards when compared to the main depression scales used in Brazil, such as BDI in its versions I and II (Cunha, 2001; Gomes-Oliveira et al., 2012). Also, the LR+ and LD− values suggest that the versions are adequate to classify cases with and without the disorder, although there is some gain for excluding cases with no diagnostic. Furthermore, the PP+ and PP− indices also reached satisfactory levels both for the full version and the short version of EBADEP-A, in all cases above 80% of probability of correct identification by the instruments.

The cut-off for the full version (48 points) and the short version (19 points) is suitable for studies and assessments that require reliable screening. However, a diagnostic interview is imperative to assess more qualitative characteristics of the depressive symptoms, since the greater the number of methods used in diagnostic, the

smaller the chances of error (Baptista & Gomes, 2011; Baptista, Gomes & Carneiro, 2013).

Another important characteristic of the two versions is that, as expected in screening scales, sensitivity should be higher than specificity, exactly what occurred in the present research (Andreoli, et al., 1998; Fletcher et al., 1996). In this sense, the practitioner who has more time to spend in the assessment/follow-up of cases and is more concerned both with quantitative and qualitative data can benefit more from using the full version of the scale, while the evaluator who needs a brief testing may benefit from the short version of EBADEP-A, since there is only a little loss of diagnostic accuracy between the versions. As it can be observed, in the full version, one case diagnosed with depression is considered as healthy (i.e., false negative), and four cases are considered as positive for depression when in fact they are not (i.e., false positive); in the short version, the numbers only rise to two and five, respectively, which suggests relatively few information is lost in this screening version.

According to the findings of this study, EBADEP-A presented adequate sensitivity, specificity, LR+, and LR− indices to detect cases with and with no diagnostic of depression (true positives and true negatives, respectively). These results are in line with the previous literature on the instrument, which demonstrates its psychometric adequacy for the evaluation of depressive symptomatology (Baptista, 2012; Bighetti, Alves, & Baptista, 2014; Souza, Baptista & Silva, 2015). Moreover, EBADEP-A might also serve an initial indicator of the presence or absence of depressive disorder. Similarly, a short version of the instrument is presented in this study for the first time, which provided similar indices of diagnostic accuracy, suggesting its adequacy for screenings in which the professional has little time to perform assistance and/or does not need further information regarding the symptoms of depression, since short versions can be very useful in screening assessments (Bolsoni & Zuardi, 2015). It can be concluded that the short version of EBADEP-A was equivalent to its full version, that is, not inferior to it. This result strongly suggests the use of the short version for screening cases, in detriment to the use of the longer version.

It is noteworthy that no studies with scales were found in the area of mental health using the equivalence concept, which is typical for comparison between procedures in other health areas (e.g., medicine and pharmacy). In this case, this study proposes the use of this concept to compare assessment tools in mental health that assess the same constructs and/or the same symptoms. Also, there are also few publications with evaluation instruments in psychiatry and, especially, in psychology, that deal with diagnostic accuracy. The application of these two concepts together aims to obtain data on the instruments that have greater practical application in comparison to studies that focus only on certain psychometric properties. Based on this, and considering the results found, it seems that this proposal can not only assist professionals with a focus on psychometric parameters, but it may also have an impact on the clinical application of the tools.

The present research presents some limitations that should be mentioned. The first one is the small sample, so that future studies should seek to replicate these findings in larger or similar samples, to increase information on the diagnostic accuracy of EBADEP-A in its full and short versions. In addition, it is interesting that studies use data collection designs randomizing patients and obtaining a clinical sample that proportionally reflects the populational prevalence of the disorder, allowing the calculation of positive and negative predictive values. It should also be considered that assessment methods in this study comprised only self-report for the target instrument, and interview for the standard gold test.

References

Akena, D., Joska, J., Obuku., E.A., Amos, T., Musisi, S., & Stein, D.J. (2012). Comparing the accuracy of brief versus long depression screening instruments which have been validated in low and middle income countries: A systematic review. *BMC Psychiatry, 12*(1), 187-199. doi: 10.1186/1471-244X-12-187

American Psychological Association. (2014). *Manual Diagnóstico e Estatístico de Transtornos Mentais– DSM-5.* Porto Alegre: ArtMed.

Andreoli, S. B., Blay, S. L. & Mari, J. J. (1998). Escalas de rastreamento de psicopatologia. *Revista de Psiquiatria Clínica, 25*(5), 229-232. Retirado de http://bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis&src=google&base=LILACS&lang=p&nextAction=lnk&exprSearch=228048&indexSearch=ID

Baptista, M. N., Gomes, J. O. & Carneiro, A. M. (2013). Exploratory Study of the Diagnostic Abilities of the Baptista Depression Scale – Adult Version (EBADEP-A). *Paidéia, 23*(56), 301-309. doi:10.1590/1982-43272356201304

Baptista, M. N. & Gomes, J. O. (2011). Escala Baptista de Depressão (Versão Adulto) – EBADEP-A: evidências de validade de construto e de critério. *Psico-USF, 16*(2), 151-161. doi: 10.1590/S1413-82712011000200004.

Baptista, M. N. (2012) *Manual técnico da Escala Baptista de Depressão em Adultos (EBADEP-A).* São Paulo: Vetor.

Baron, E. C., Davies, T., & Lund, C. (2017). Validation of the 10-item Centre for Epidemiological Studies Depression Scale (CES-D-10) in Zulu, Xhosa and Afrikaans populations in South Africa. *BMC Psychiatry, 17(*1), 1-12. doi: 10.1186/s12888-016-1178-x

Bech, P., Wilson, P., Wessel, T., et al. (2009). A validation analysis of two self-reported HAM-D6 versions. *Acta Psychiatrica Scandinavica, 119*(4), 298-303. doi: 10.1111/j.1600-0447.2008.01289.x

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1997). *Terapia cognitiva da depressão* (S. Costa, Trad.). Porto Alegre: Artmed.

Bighetti, C. A., Alves, G. A. S., & Baptista, M. N. (2014). Escala Baptista de Depressão (EBADEP-A): evidências de validade com o Big Five. *Avaliação Psicológica, 13*(1), 29-36. Recuperado de http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712014000100005&lng=pt&tlng=pt.

Björgvinsson, T., Kertz, S. J., Bigda-Peyton, J. S., McCoy, K. L., & Aderka, I. M. (2014). Psychometric Properties of the CES-D-10 in a Psychiatric Sample. *Assessment, 20*(4), 429-436. doi: 10.1177/1073191113481998

Bolsoni, L. M., & Zuardi, A. W. (2015). Estudos psicométricos de instrumentos breves de rastreio para múltiplos transtornos mentais. *Jornal Brasileiro de Psiquiatria, 64*(1), 63-69. doi: 10.1590/0047-2085000000058

Bromet, E., Andrade, L. H., Hwang, I., Sampson, N. I., Alonso, J., Girolamo, G., ... & Karam, A. N. (2011). Cross-national epidemiology os DSM-IV major depressive episode. *BMC Medicine, 9*(1), 1-16. doi: 10.1186/1741-7015-9-90

Calil, H. M., & Pires, M. L. N. (1998). Aspectos gerais das escalas de avaliação de depressão. *Revista de Psiquiatria Clínica, 25*(5), 240-244. doi: 11.1178/02258358849637132008

Conselho Federal de Psicologia. Resolução n° 002, de 24 de março de 2003: Define e regulamenta o uso, a elaboração e a comercialização de testes psicológicos e revoga a Resolução CFP n° 025/2001. 2003. Retirado de http://site.cfp.org.br/resolucoes/resolucao-n-2-2003/

Cunha, J. (2001). *Manual em português das escalas Beck.* São Paulo: Casa do Psicólogo.

Del-Ben, C. M., Vilela, J. A. A., Crippa, J. A., Hallak, J. E., Labate, C. M., & Zuardi, A. W. (2001). Confiabilidade da "Entrevista Clínica Estruturada para o DSM-IV - Versão Clínica" traduzida para o português. *Revista Brasileira de Psiquiatria, 23*(3), 156-159. doi: 10.1590/S1516-44462001000300008

Derogatis, L.R. & Lynn, L.L. (2000). Psychological tests in screening for psychiatric disorder. In Maruish, M. E. The *Use of Psychological Testing for Treatment Planning and Outcomes Assessment*. New Jersey: Lawrence Erlbaum. doi: 10.1023/A:1005515421103

Ferster, C. B., Culbertson, S., & Boren, M. C. (1977). Depressão clínica. In C. B. Ferster, S. Culbertson, & M. C. Boren (Eds.), *Princípios do comportamento* (pp. 699-725). São Paulo: Hucitec.

Fletcher, R. H, Fletcher, S. W. & Wagner, E. H. (1996). *Epidemiologia clínica: Elementos essenciais.* Porto Alegre: Artmed.

Gomes, J. O. & Baptista, M. N. (2014). Normalization procedure for the Baptista Depression Scale - Adult Version (EBADEP-A): transferring of norms. *Avances en Psicología Latinoamericana, 32*(3), 419-432. doi: 10.12804/apl32.03.2014.02.

Gomes-Oliveira, M. H., Gorenstein, C., Lotufo-Neto, F., Andrade, L. H. & Wang, Y. P. (2012). Validation of the Brazilian Portuguese Version of the Beck Depression Inventory-II in a community sample. *Revista Brasileira de Psiquiatria, 34*(4), 389-394. doi:10.1187/5047-89722056321477

Goreinstein, C., Wang, Y. P., Argimon, I. L. & Werlang, B. S. G. (2011). *Manual do Inventário de Depressão de Beck – BDI-II.* São Paulo: Casa do Psicólogo.

Greene, C. J., Morland, L. A., Durkalski, V. L., & Frueh, B. C. (2008). Noninferiority and equivalence designs: issues and implications for mental health research. *Journal of traumatic stress, 21*(5), 433-439. doi: 10.1002/jts.20367

Hajian-Tilaki, K., 2013. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine, 4*(2), 627-635.

Hanson, L. L. M., Westerlund, H., Leineweber, C., Rugulies, R., Osika, W., Theorell, T., & Bech P. (2014). The Symptom Checklist-core depression (SCL-CD6) scale: Psychometric properties of a brief six item scale for the assessment of depression. *Scandinavian Journal of Public Health, 42*(1):82-88doi:10.1590/S0034-89102012005000044

Iared, W., & Valente, O. (2009). Revisões sistemáticas de estudos de acurácia. *Diagnóstico & Tratamento, 14*(2), 85-8. doi: 10.1278/0114545888963255479

Klein, C. H. & Costa, E. A. (1987). Os erros de classificação e os resultados de estudos epidemiológicos. *Cadernos de saúde pública, 3*(3), 35-46. doi: 10.1023/A:1005515421103

Kroenke, K. (2001). Depression screening is not enough. *Annals of Internal Medicine, 134*(5), 418–420. doi:10.1187/5047-89722056321477

Kullback, S. (1959). *Information theory and statistics.* New York: John Wiley & Sons.

Lesaffre, E. (2008). Superiority, Equivalence, and Non-Inferiority Trials. *Bulletin of the NYU Hospital for Joint Diseases, 66*(2),150-154.

Mitchell, A. J., Vaze, A., & Rao, S. (2009) Clinical diagnosis of depression in primary care: A meta-analysis. *The Lancet, 374*(9690), 609-619. doi:10.1187/5047-89722056321477

Mitchell, A. J., & Coyne, J. C. (2007) Do ultra-short screening instruments accurately detect depression in primary care? *British Journal of General Practice, 57*(535), 144-151. doi: 10.1023/A:10055785422503

Obuchowski, N. A. (2000). Sample size tables for Receiver Operating Characteristic studies. *American Journal of Roentgenology, 175*(3), 603-608. doi: 10.2214/ajr.175.3.1750603.

Parshall, M. B. (2013). Unpacking the 2x2 table. *Heart & Lung, 42*(3), 221-226. doi: 10.1016/j.hrtlng.2013.01.006

Pirraglia, P. A., Rosen, A. B., Hermann, R. C., Olchanski, N. V., & Neumann P. (2004). Cost-utility analysis studies of depression management: A systematic review. *American Journal of Psychiatry, 161*(12), 2155–2162. doi: 10.1590/S1413- 81232009000200007

Santor, D. A., Gregus, M., & Welch, A. (2006). Eight Decades of Measurement in Depression. *Measurement, 4*(3), 135-155. doi: 10.1590/S1413- 89123200900028743 07

Soares, O. T., Moreno, D. H., Moura, E. C., Angst, J., & Moreno, R. A. (2010). Confiabilidade e validação da versão brasileira do Questionário de Hipomania (HCL-32 VB) comparado ao Questionário de Transtornos de Humor (MDQ). *Revista Brasileira de Psiquiatria, 32*(4), 416-423. doi: 10.1590/S1516-44462010000400015.

Souza, M. S., Baptista, M. N., & Silva, G. A. (2015). Estudos psicométricos preliminares da Escala Baptista de Depressão para Adultos. *Estudos de Psicologia, 32*(3), 357-370. doi.org/10.1590/S1413-82712012000300007

Streiner, D.L. (2003). Diagnosing tests: using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81*(3), 209-219. doi: 10.1207/S15327752JPA8103_03

World Health Organization (1993). *International Classification of Diseases – ICD-10.* Geneva: WHO.
Zhou, X. H., Obuchowski, N.A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine.* New York: Wiley.

recebido em setembro de 2017
aceito em agosto de 2018

Sobre os autores

Makilim Nunes Baptista é psicólogo, Doutor pelo departamento de Psiquiatria e Psicologia Médica da Universidade Federal de São Paulo, docente do Programa de Pós-Graduação *Stricto Sensu* em Psicologia da Universidade São Francisco, Campinas/SP e pesquisador bolsista produtividade pelo CNPq.

Lucas de Francisco Carvalho é psicólogo e Doutor em Psicologia com ênfase em Avaliação Psicológica pela Universidade São Francisco. Atualmente é docente do Programa de Pós-Graduação *Stricto Sensu* em Psicologia da Universidade São Francisco, Campinas/SP e pesquisador bolsista produtividade pelo CNPq.