

ANNEX

English version of article “*Outcome Questionnaire (OQ-45.2): avaliação das propriedades psicométricas via modelo bifactor e TRI*” published in journal *Psico* (v. 47, n. 4, p. 298-308) <<http://revistaseletronicas.pucrs.br/ojs/index.php/revistapsico>>.

Outcome Questionnaire (OQ-45.2): assessment of the psychometric properties using bifactor model and IRT

Sonia Maria da Silva

Universidade Guarulhos, SP, Brasil

Iraí Cristina Boccato Alves

Instituto de Psicologia USP, SP, Brasil

Evandro Moraes Peixoto

Universidade São Francisco, SP, Brasil

Glaucia Mitsuko Ataka Rocha

Instituto de Psicologia USP, SP, Brasil

Tatiana de Cássia Nakano

Pontifícia Universidade Católica de Campinas, SP, Brasil

Abstract

The aim of this study was to evaluate the psychometric properties of the Brazilian version of the Outcome Questionnaire – OQ-45.2, an instrument that assesses the progress of patients undergoing psychotherapy. The adjustments of different measurement models proposed to OQ-45.2 were compared through Confirmatory Factor Analysis, and the parameters of items and participants were estimated using Andrich's rating scale model. The sample comprised 419 adults (mean age: 32.18±14.3; 62.8% female). The results demonstrated the suitability of the bifactor model with three specific factors (symptom distress, interpersonal relationships and social roles) and a general factor (overall maladjustment) when compared to other models in the literature, such as the one-factor and the three related factors models. The invariance of the OQ-45's internal structure was observed, evaluating both male and females. With regard to the items' properties (adjustment and difficulty) the appropriate psychometric parameters were obtained for the assessment of the psychotherapeutic outcome. In conclusion, the OQ-45.2 is a suitable measurement tool for assessing these characteristics in the Brazilian population.

Keywords: Item Response Theory; Factorial Analysis; Psychological assessment; Test validity; Psychotherapy.

Outcome Questionnaire (OQ-45.2): avaliação das propriedades psicométricas via modelo bifactor e TRI

Resumo

O objetivo foi avaliar as propriedades psicométricas da versão brasileira do *Outcome Questionnaire (OQ-45.2)*, instrumento designado a avaliar o progresso de pacientes em psicoterapia. Comparou-se o ajuste de diferentes modelos de medidas propostos ao *OQ-45.2* através da Análise Fatorial Confirmatória e verificou-se os parâmetros dos itens e dos participantes por meio do Modelo de Resposta Gradual. 419 adultos responderam ao instrumento (32,18±14,3; 62,8%, mulheres). Os resultados demonstram a adequação da estrutura *bifactor* composta por três fatores específicos (desconforto subjetivo, relações interpessoais e desempenho do papel social) e um fator geral (desajustamento global) quando comparados aos outros modelos: unifatorial e com três fatores correlacionados. Verificou-se a invariância da estrutura interna do *OQ-45.2* ao avaliar homens e mulheres. Em relação às propriedades dos itens (dificuldade e ajustamento) foram obtidos parâmetros psicométricos apropriados para a avaliação de resultados psicoterapêuticos. Conclui-se que o *OQ-45.2* é uma medida adequada para avaliação destas características na população brasileira.

Palavras-chave: Teoria de Resposta ao Item; Análise Fatorial; Avaliação psicológica; Validade do teste; Psicoterapia.

Outcome Questionnaire (OQ-45.2): evaluación de las propiedades psicométricas via modelo bifactor y TRI

Resumen

El objetivo fue evaluar las propiedades psicométricas de la versión brasileña del *Outcome-Questionnaire (OQ-45.2)*, instrumento diseñado para evaluar el progreso de los pacientes en psicoterapia. Ha sido comparado el ajuste de los diferentes modelos de medidas para *OQ-45.2* mediante análisis factorial confirmatorio y ha sido encontrado los parámetros de los ítems y de los participantes a través del Modelo de Respuesta Gradual. 419 adultos respondieron el instrumento (32,18±14,3; 62,8% mujeres). Los resultados demuestran la adecuación del modelo *bifactor* formado por tres factores específicos (el malestar subjetivo, las relaciones interpersonales y el desempeño del papel social) y un general (desajuste global) cuando se compara con otros modelos de la literatura: un factor de tres y factores relacionados. Se observó la invariancia de la estructura interna del *OQ-45.2* cuando se evaluaron personas del sexo femenino y masculino. Con respecto a las propiedades de los elementos (ajuste de dificultad) se obtuvieron parámetros psicométricos apropiados para la evaluación de los resultados psicoterapêuticos. Se concluyó que el instrumento presenta buenas propiedades psicométricas para evaluación de la población brasileña.

Palabras clave: Teoría de Respuesta al Item; Análisis factorial; Evaluación psicológica; Validación de test; Psicoterapia.

In recent years, Brazil, has seen increasing levels of research into the process and results of psychotherapy (Brum et al., 2012; Del Prette, 2011; Pieta, Siegmund, Gomes, & Gauer, 2015), particularly in university clinics, which has led to a need to satisfy three objectives: provision of community services, professional training for students in Psychology courses and the development of research studies. The demands of users require constant “thinking” and “doing” of a Psychology that responds to these needs, combining theory and practice by means of the production of knowledge, particularly the development and/or adaptation of psychological instruments that are suitable for the population served and which make it possible to measure change in psychotherapeutic processes (Honda, Peixoto, Rocha, & Enéas, 2015).

There is a concern to seek evaluative practices based on empirical evidence (Antony & Barlow, 2010). Psychometric properties, such as validity and reliability, have become hot topics of conversation, particularly with the publication of Brazil’s Federal Psychology Council Resolution 002/2003, which considers the need to refine the tools and technical procedures of the work of psychologists as well as to perform a periodic review thereof.

Within the range of psychological evaluation tools available, those referred to as self-reporting have been gaining prominence since, as they are answered by the individuals themselves, they permit an assessment of the way in which subjects report the perception of their own condition. These tools are being increasingly used in the area of evaluation of the psychotherapy process and results, as these measures appear to be making a more significant contribution to determining changes in patients (Harmon, Hawkins, Lambert, Slade, & Whipple, 2005). They are simple scales which entail less effort and which succeed in drawing research and clinical practice closer together (Lambert, Hansen, & Finch, 2001).

Serralta, Nunes and Eizirik (2007), stressed that, in Brazil, there was little systemization in terms of the use of instruments to evaluate psychotherapeutic processes, as well as in relation to the forms of assistance and the results of psychotherapy. The same authors add that psychotherapy research is still in its early development, if compared to international research, and in-depth, systematic studies are still lacking on this topic. To help fill this gap, the main objective of this study was to assess proof of validity based on the internal structure and reliability of the Outcome Questionnaire (OQ-45.2) (Lambert et al., 2004) for the Brazilian population.

The OQ-45.2 is one of the most popular instruments among North American professionals that aim to

evaluate the outcome of psychotherapeutic processes (Hatfield & Ogles, 2004). This popularity can also be seen in the number of languages into which the tool has been translated, namely Japanese, Korean, French, Italian, German, Dutch, European Portuguese, Spanish, Swedish, Norwegian and Russian (Huag, Puschner, Lambert, & Kordy, 2004; Jong et al., 2007; Lambert et al., 2004, Machado & Fassnacht, 2015).

Developed by Lambert, Lunnen, Umphress, Hansen and Burlingame in 1994 (Lambert & Finch, 1999), the OQ-45 was designed for use on three distinct application levels: 1) to measure the current dimension of psychic suffering; 2) to evaluate results both before and after interventions in treatment or to monitor the patient’s response to treatment; 3) to monitor the psychotherapist’s decision-making process by means of a standardized instrument, in order to improve quality of treatment (Lambert et al., 2004). The author reports that the instrument was not designed for patient diagnosis as this task can be carried out with instruments developed specifically for this purpose, such as the MMPI-2 in the USA, which require longer application times.

According to the measurement model proposed by Lambert (1983), the progress of the patient undergoing psychotherapy should be evaluated through three aspects of their lives: symptom distress, interpersonal relationships and the performance of the social role. These areas of functioning suggest a continuum of sentiments and perceived sensations concerning their internal state, what they are experiencing privately, how they relate to significant people in their lives and how to deal with tasks related to productivity, for example, work, school or any other activity, including leisure activity. The development and selection of items for the OQ-45 were determined based on a series of considerations. Initially, questions were selected related to dysfunctions or the more common problems with a wide variety of disorders. Then comes a list of items that address those symptoms that have a greater probability of occurrence, regardless of the peculiarities of the problem. Then, in the third stage, items evaluating relevant personal and social characteristics in terms of quality of life. Based on these considerations, the items were grouped theoretically into three subscales referred to as Symptom Distress (SD), Interpersonal Relationships (IR) and Social Role (SR) (Lambert et al., 2004).

Diverse studies have been carried out in different countries with the aim of evaluating the psychometric properties of the OQ-45. Of these, the more recent studies have found good indicators of reliability, both for the subscales and for the total score. The study

by Boswell, White, Sims, Harrist, & Romans (2013), evaluating a sample of 220 university students in the USA, reported Cronbach's α for $SD=.93$; $IR=.78$; $SR=.70$ and a Total Score $=.94$. A study using the German version of the scale applied to psychiatric patients ($N=294$) (Puschner, Cosh, & Becker, 2015) also observed similar coefficients (Cronbach's α for $SD=.93$; $IR=.74$; $SR=.68$ and a Total Score $=.93$), as well as the study by Machado & Fassnacht (2015) conducted using different strata of the Portuguese population, in which coefficients between .61 and .92 were observed in samples of university students, .59 and .92 in a sample of the general public and .56 and .93 for the clinical sample.

The version adapted for Brazilian Portuguese by Carvalho and Rocha (2009) was evaluated by Silva (2013), who found Cronbach's α for $SD=.91$ in the non-clinical sample and .9 in the clinical sample; $IR=.71$ and .64; $SR=.65$ and .68; Total Score $=.93$ and .92 in the respective samples. Moreover, the author performed a study of the reliability of the Brazilian version, through the test and retest method, on a sample of university students ($n=33$) and clinical sample, on patients in a university clinic ($n=55$), the application interval ranging from seven to 14 days. The results revealed Pearson correlations between the different applications varying from .58 to .91 for the university students, and .75 to .89 for the sample of patients.

If, on the one hand, the reliability of the scale is frequently evaluated as adequate, on the other, some studies have encountered difficulty in providing empirical support for the three-factor structure. Jong et al. (2007), in a study with the Dutch version, using Exploratory Factor Analysis (EFA), found two additional factors: one comprising social role items and the other reflecting anxiety and somatic symptoms. Mueller, Lambert and Burlingame (1998), on the other hand, through Confirmatory Factor Analysis (CFA), detected better indices of adjustment for the unifactorial structure versus the three-factor structure. More recently, studies have found proof of validity in a bifactor structure for the tool in question. In this model, it is assumed that the items on the scale capture two different sources of variance, specific variance, arising from three specific factors, and shared common variance among all the items, also called a general factor or Overall Maladjustment (Lambert et al., 2004; Lo Coco, et al., 2008).

The use of psychological instruments in different countries, adapted for languages other than the original, require an equivalent version, as the underlying psychological constructs need to be equivalent

(Butcher, Derksen, Sloore, & Sirigatti, 2003). In order to be used in our country, Carvalho and Rocha (2009) translated and adapted the OQ-45.2 based on the original version and on the Portuguese version, having produced a Brazilian version for the development of future studies on psychometric properties.

Thus, it is necessary to conduct studies to assess the proof of validity based on the instrument's internal structure for this population, given that the absence of such information compromises any inference based on the test scores (AERA, APA, & NCME, 2014). Another significant gap in respect of this instrument relates to the lack of information in terms of its ability to evaluate similarly, people of different sexes. This gap is all the more evident inasmuch as the participants of the studies are generally compared through the OQ-45.2 scores based on this variable (Lambert et al., 2004; Jong et al., 2007; Machado & Fassnacht, 2015; Rodríguez, 2000). Accordingly, getting empirical evidence that the observed variables, the scale items, are related to the latent constructs in the same way for the different groups, has become an imperative for researchers in the area, as this is a prerequisite for comparing these groups via the raw results derived from the scale (Milfont & Fisher, 2010).

In order to provide contributions to fill these gaps, the main aim of the present study was a) to assess the initial validity evidence based on internal structure and reliability of the OQ-45.2 for the Brazilian population, b) to evaluate the invariance of the internal structure of this measure with regard to the sex of the participants, and c) to assess the parameters of the items (difficulty and adjustment) and the participants (level of intensity in the construct).

Method

Participants

The sample obtained was made up of 419 participants, of which 263 (62.8%) were female, with ages varying from 18 to 78 ($M=32.87 \pm 15.6$). As far as the level of schooling is concerned, of the total sample, 46% of the participants had entered higher education courses, 39% had attended high school, 8% had attended primary education and 7% did not respond. The total sample ($N=419$) was subdivided into two groups: patients ($N=59$), patients being treated in a Psychology university clinic situated in a large city in Greater São Paulo, and non-patients ($N=360$). As for age grouping, 59% were aged between 20 and 39, 21% between 40 and 59, 12% between 18 and 19, 7% between 60 and 78 and 1% did not respond or did not give their age.

The sample was also described in terms of economic resources. For this classification, the Brazil Economic Classification Criterion published by the Brazilian Market Research Association (*ABEP*) was used, a criterion whose function it is to assess individuals' purchasing power. *ABEP*'s classification of economic classes is divided into eight parts, each one of which corresponds to an average family income (gross value in Brazilian Real) A1=12,926; A2=8,418; B1=4,418; B2=2,565; C1=1,541; C2=1,024; D=714 and E=477. In the *ABEP* classification, based on the 2010 Socioeconomic Survey, 35% are classified in Class B2; 25.5% in C1; 13% in C2; 11% in B1; 5% in A2; 3% in D; .5% in E and 7% did not respond.

Instruments

Outcome Evaluation Scale OQ-45.2 (Lambert et al., 1994). This is a self-reporting tool comprising 45 items, whose responses are given on a Likert-type five-point scale that ranges from "never" (zero points) through "rarely" (one point), "sometimes" (two points), "frequently" (three points) to "almost always" (four points). According to the original study (Lambert, 1996), the tool comprises three factors: Symptom Distress (25 items), Interpersonal relationships (11 items), and Social role (9 items). More recent studies, however, have presented evidence of a bifactor structure for the instrument in question (Lambert et al., 2004; Lo Coco et al., 2008); in this perspective, the internal structure is composed of a general factor known as Overall Maladjustment and the abovementioned three specific factors. The version used for this study is the result of a translation and cultural adaptation of the OQ-45.2 conducted by Carvalho and Rocha (2009), following authorization by the American authors for the translation, cultural adaptation and semantic equivalence for use in Brazil.

Sociodemographic questionnaire: instrument developed to access the main characteristics of the sample such as sex, age, level of schooling and socioeconomic level.

Ethical considerations

The Design for this Research Study was approved by the university's Ethics Committee under protocol reference number CAAE 0041.0.132.000-110. The tools were applied to the participants, each of whom received a Free and Informed Consent Form which included the objectives of the study and the form of disclosure of the results, in accordance with ethical standards. Only those people agreeing to the study's procedures and signing the consent form could participate in the study. It should be stressed that, for

the participants in respect of the patients group, the instruments were applied by researchers external to the psychotherapeutic process.

Procedures and Statistical Analyses

In order to obtain proof of validity of the internal structure of the OQ-45.2, the Confirmatory Factor Analysis (CFA) was employed. Thus, using the χ^2 difference test, the following models were tested and compared: one-factor, three-factor and bifactor solution. To this end, the Robust Weighted Least Square (WLSMV) estimation method was used. The choice of these procedures was based on the fit of these methods to the level of ordinal measurement (Lara & Alexis, 2014). In this regard, the models were tested based on indices recommended by Muthén and Muthén (2012): WLSMV χ^2 , gl, CFI, TLI and RMSEA. These analyses were conducted with the support of the statistical package Mplus 7.11 (Muthén and Muthén, 2012). Once validity evidence based on internal structure was obtained, the invariance of the measurement model was evaluated between participants of the female and male sexes (Milfont & Fisher, 2010).

Still on the topic of the internal structure, the model was checked to see if it presented an essential unidimensionality within the bifactor structure. To this end, the procedures described by Rios and Wells (2014) were employed, namely the Expected common variance (ECV) and Percentage of uncontaminated correlations (PUC). According to the authors, the former procedure provides the ratio between the power of the general factor and the specific factors in the bifactor model. ECV values close to 1 prove that there is a strong overall factor present in the bifactor data. Meanwhile the PUC may be defined as the number of uncontaminated correlations divided by the number of unique correlations. PUC values close to 1 prove that a strong overall factor is present in the bifactor data. The last step consists of an evaluation of the degree to which the total score reflects a common variable. To this end, an evaluation of reliability was employed checking the effect of the general factor via the procedure referred to by Reise (2012) as coefficient omega (hierarchical) ω_H .

For a description of the item and participant parameters, the Item Response Theory (IRT), or more specifically Andrich's rating scale model was employed, with a calibration estimated based on the maximum likelihood method, available in the Winsteps (Linacre, 2015) software application. In order to identify the scale metrics, the mean of the indices of difficulty of the items was anchored at 0. In

this way, different statistical procedures were used to obtain the latent trait level exhibited by the subjects (*theta*), the indices of difficulty of the items (*b*), the indices of adjustment of the items (infit and outfit), the response characteristics curve and the indices of reliability.

Results

Table 1 displays the indices of adjustment in the one-factor, three-factor and bifactor models.

According to Table 1, it can be seen that the bifactor model has significantly higher indices of adjustment than the other models. According to Rios and Wells (2014), further evidence that the data fit the bifactor structure is derived from the comparison between the bifactor model and unifactorial model by means of changes in the CFI index (ΔCFI), as the former model has a ΔCFI value greater than .01. The factor loading of the items in the bifactor model are shown in Table 2.

With regard to the factor loading shown for the items, Table 2 shows that, in generally terms, these are higher for the general factor when comparing the factor loading presented by the specific factors, which denotes a higher general factor variance in the proposed structure. Nevertheless, some exceptions were observed, namely with items 7, 17, 27 and 38. Notably some items had high negative loads, greater than or equal to .30, in the specific factors, namely items 1 and 30. The fact that items 11, 14 and 26 do not have factor loading greater than or equal to .3, neither for the overall factor nor for the respective specific factor, should be highlighted.

Given that the respective ECV and PUC values were .75 and .61, this suggests the presence of essential unidimensionality in the model in question. Results complementary to those obtained in the coefficient omega procedure: equal to .92 for the general factor and .02, .00004 and .01, for the

respective specific factors, suggest that a very high variance in the scores could be attributable to the general factor.

As for the evaluation of invariance in the model, the results indicate an equivalence of the configural model between the groups, ($\chi^2/df=1.39$; CFI=.938; TLI=.935; RMSEA=.044 CI 90% .039-.048), i.e. similarities in the overall internal structure proposed for the measurement model, the number of latent variables. As with the invariance of the scale model which, in addition to the equivalence described in the configural model, evaluates if the mean of the items is equivalent between the groups, having checked possible differences to latent variables, is equivalent for both groups ($\chi^2/df=1.32$; CFI=.941; TLI=.942; RMSEA=.040 CI 90% = .035-.044), considering that the ΔCFI and ΔTLI between the two models are lower than .01. More specifically, these results show that men and women of the same average in the latent variable do not present different means in the items. According to Santos and Primi (2014), the scale model test is equivalent to the Item Differential Functioning (DIF) analysis in the difficulty parameter “*b*” in the IRT. Therefore, it may be inferred that the tool is capable of evaluating men and women in similar fashion, and that the possible differences that exist between the raw test scores, when comparing these groups, could be linked to the psychological characteristics of the subjects being evaluated, and not to measurement error associated with the tool.

With regard to the participant and item parameters estimated using IRT, the results are displayed in Table 3, where the indices of difficulty of the items can be found, along with the indices of adjustment of the items (infit and outfit), indices of correlation between item and theta presented by the participants, and the descriptive statistics of the participant parameter (*theta*).

As depicted in Table 3, the analyses were conducted considering the general factor evaluated through the

TABLE 1
Indices of confirmatory adjustments of the one-factor, three factors and bifactor models of the OQ-45

Model evaluated	χ^2	Df	χ^2/df	$\Delta \chi^2$	Δdf	CFI	TLI	RMSEA	RMSEA CI 90%
Unifactorial	1960.298	945	2.07			.90	.89	.051	.048-.054
Three factors	1862.425	942	1.98	97.973 <i>p</i> <.001	3	.91	.90	.049	.045-.052
Bifactor	1665.158	900	1.85	197.267 <i>p</i> <.001	42	.93	.92	.045	.042-.049

TABLE 2
Factorial loads of the items

Items	General Factor	Specific Factors			θ^2
		1	2	3	
2	.308	.220			.857
3	.308	.123			.890
5	.613	.027			.624
6	.618	.176			.587
8	.756	.033			.427
9	.765	-.036			.413
10	.674	.082			.539
11	.177	.233			.914
13	.711	-.223			.444
15	.790	-.099			.366
22	.464	.207			.742
23	.673	.140			.527
24	.656	-.184			.535
25	.646	.272			.509
27	.307	.387			.756
29	.538	.360			.581
31	.561	-.209			.642
33	.645	.070			.579
34	.423	.233			.767
35	.462	.242			.728
36	.653	.311			.477
40	.716	.346			.367
41	.595	.282			.566
42	.777	.141			.377
45	.386	.309			.756
1	.370		-.423		.684
7	.368		.396		.708
16	.337		-.017		.886
17	.403		.488		.599
18	.688		.087		.519
19	.472		-.034		.776
20	.612		.152		.603
26	.255		-.094		.926
30	.528		-.316		.621
37	.521		.320		.626
43	.582		-.058		.658
4	.404			.281	.758
12	.362			.314	.770
14	-.024			-.046	.997
21	.502			-.055	.744
28	.496			.431	.568
32	.271			.304	.834
38	.438			.554	.501
39	.392			.405	.682
44	.586			.409	.488
$\Sigma\lambda^2$	12.92	1.2	0.82	1.10	
$\Sigma\lambda$	518.9284	11.8336	0.25	6.76	

$\Sigma\lambda^2$ = sum of the squares of the factor loading; $\Sigma\lambda$ = sum of the factor loading squared; θ^2 = residual variance of item.

TABLE 3
Descriptive statistics of the indices of difficulty and adjustment of the items

<i>Items</i>	<i>b</i>	<i>S.E.</i>	<i>Infit</i>	<i>Outfit</i>	<i>r</i>	<i>Theta</i>
1	.55	.07	1.11	1.24	.39	
2	-.52	.06	1.09	1.07	.35	
3	-.28	.05	1.23	1.40	.32	
4	-.66	.05	1.06	1.08	.45	
5	-.37	.05	.88	.86	.59	
6	-.57	.05	.84	.84	.60	
7	-.24	.04	1.21	1.42	.36	
8	.71	.07	.78	.69	.53	
9	.12	.05	.75	.76	.66	
10	.08	.05	.83	.78	.61	
11	.80	.09	1.48	2.88	.14	
12	-.16	.05	1.18	1.19	.37	
13	.19	.05	.83	.91	.58	
14	-1.15	.05	1.64	1.96	.05	
15	.17	.05	.73	.70	.67	
16	-.90	.05	1.23	1.28	.36	
17	-.05	.04	1.27	1.43	.38	
18	-.26	.04	.85	.92	.61	
19	-.13	.05	1.03	1.02	.45	
20	.13	.05	.91	.89	.55	
21	-.19	.05	1.08	1.09	.47	
22	-.58	.05	1.00	.99	.48	
23	.08	.05	.84	.86	.60	
24	.23	.05	1.01	1.15	.49	
25	-.09	.05	.86	.87	.60	
26	.88	.08	1.39	1.53	.19	
27	.10	.05	1.21	1.29	.36	
28	-.04	.05	1.00	1.04	.51	
29	.17	.05	.92	.91	.54	
30	.35	.05	1.01	1.04	.47	
31	-.01	.05	1.07	1.34	.45	
32	1.42	.13	1.10	1.42	.15	
33	.24	.05	.86	.82	.58	
34	-.18	.05	1.08	1.11	.45	
35	.33	.05	1.14	1.22	.41	
36	-.28	.05	.77	.77	.66	
37	-.48	.04	1.07	1.12	.50	
38	-.10	.05	1.05	1.11	.45	
39	.47	.06	1.07	1.05	.39	
40	.12	.05	.75	.72	.65	
41	-.01	.05	.90	.94	.56	
42	-.32	.05	.69	.69	.72	
43	.01	.05	.96	1.01	.52	
44	.68	.06	.90	.82	.50	
45	-.29	.05	1.07	1.11	.44	
M	.00	.05	1.02	1.10	0.47	-0.71
SD	.47	.01	.20	.37	0.14	0.57
Min.	1.42	.04	1.64	2.88	0.05	1.12
Max.	-1.15	0.13	0.69	0.69	0.72	-2.59

b = indices of difficulty; *r* = correlation between item and theta; *theta* = latent trait level presented by the participants.

OQ-45.2, considering the indications of essential unidimensionality found through the EVC and the PUC. As far as the indices of difficulty of the items are concerned, a small variation can be seen around the mean; it should be stressed that using the anchoring system, the mean difficulty of the items was centered at zero. These results show that no item was more easy or more difficult for the participants to endorse. From another perspective, it indicates a better ability of the instrument to evaluate the central portion of the Overall Maladjustment continuum.

With regard to the Infit adjustment indices, except for items 11, 14 and 26, it can be seen that the items exhibited indices considered adequate (between .7 and 1.3), as recommended by Bond & Fox (2001). This indicates an adjustment of the items to the response patterns expected in the model when the values for difficulty of the items are close to the *theta* values of the individuals. As for the Outfit indices, notably items 7, 11, 14, 17, 26 and 32, these exhibited values outside of the established range (i.e. between .7 and 1.3), thereby indicating response patterns that are not expected by the model when the difference between *theta* and the difficulty of the categories is very large. Similarly, the indices of correlation between item and the participants' level of *theta* show the outfit of items 11, 14, 26 and 32, given the inability of these items, unlike the others in the scale, to restore the participants' level of *theta*. In terms of the participants' parameters, the descriptive statistics for *theta* indicate that the subjects tend to endorse lower response categories in the tool's items ($M = -.71$ and $SD = .57$). Using Maximum and Minimum statistics, a high variability can be seen with the participants' level of *theta* (between -2.59 and 1.12), which shows that the sample was composed of people with different levels of overall maladjustment.

In terms of the probability of the respondents endorsing each response category of the items comprising the dimensions of the OQ-45.2, a graphical analysis of the response characteristics curves indicated a growing monotonic relationship between the values of *theta* and the scale categories, i.e. between the subjects' level of ability and the level of difficulty presented by each of the response categories. Generally speaking, this demonstrates the functionality of the response categories on the five-point Likert scale used in the instrument. Lastly, the indices of reliability for the instrument were evaluated without controlling the general factor, and the results showed Cronbach alpha indices of .94 for the general factor Overall Maladjustment and indices equal to .91, .72 and .70 for the respective specific factors SD, IR and SR.

Discussion

The aim of this study was to evaluate the psychometric properties of the OQ-45 for a sample of the Brazilian population. To this end, we resorted to a verification of the indices of adjustment of three different factorial structures commonly used in the literature (Lo Coco et al., 2008) namely: one-dimensional structure, three factors and bifactor. Confirming the theoretical hypothesis, the results suggest the suitability of the bifactor structure of the OQ-45.2, composed of three specific factors: SD, IR and SR and a general factor, Overall Maladjustment (Lambert et al., 2004). It should be stressed that the three specific factors correspond to the original proposal in the scale proposed by Lambert et al. (1994).

Although this study confirms the results of earlier studies relating to the bifactor structure of the OQ-45.2, it can be observed that the indices of adjustment obtained in the CFA conducted in these studies are relatively higher than those obtained by other researchers, such as Lo Coco et al. (2008), evaluating the bifactor structure in the Italian population ($\chi^2/df = 2.99$; CFI .830; AGFI .806; SRMR .054; RMSEA .049). This evidence indicates the suitability of the estimation method employed, the WLSMV, which is based upon a polychoric correlation matrix which, therefore, is more suitable to the level of ordinal measurement, as in the case of the psychological tests responded to on the Likert-type scale (Cook, Kallen, & Amtmann, 2009; Muthén and Muthén, 2012), in contrast to the maximum likelihood estimation method used in the abovementioned study.

Staying with the topic of the internal structure, the invariance of the factorial structure stands out when compared to the evaluation of the different groups by the sex of the participants. This evidences the capacity of the scale to evaluate such groups in similar fashion, so as to enable a comparison between the scores in these groups in future studies (Cook, Kallen, & Amtmann, 2009; Milfont & Fisher, 2010). Various authors have highlighted the importance of this procedure for psychological tests given that the lack of such information could lead to researchers/professionals in practice making inappropriate comparisons, even infringing ethical issues associated with psychological evaluations (Milfont & Fisher, 2010).

Using the IRT, it was possible to verify the parameters of the items making up the scale, results which evidence the power and limitations of the component items. Among the positive points are the infit and outfit indices which, for the majority of the items, fit the response pattern expected by the model, as

well as the suitability of the Likert-type scale adopted in the tool. As for the weaker points of the scale, four items may be identified, namely 11, 14, 26 and 32, with indices of adjustment beyond those expected, as well as a low capacity to restore the *theta* of the participants.

According to Wright & Linacre, (1994) the Outfit statistic is based on unexpected peripheral values and, therefore, is more sensitive to outlier influence. Accordingly, it may be inferred that these are less troublesome situations with regard to the indices of adjustment of the items, given the greater importance of the item's capacity to fit the response pattern expected by the model, when the subjects' latent trait levels are close to the level of intensity of the items (Infit). This explains why item 32 has been retained.

Lastly, the analyses suggest that items 11, 14 and 26 be excluded from the scale, as these did not present factorial loads greater than or equal to .30 for the specific factors, nor for the general factor, as theoretically expected. These indications were confirmed using the IRT, as the items in question did not exhibit good indices of adjustment or correlation with the estimate of the subjects' latent trait level. It should be emphasized that the incompatibility of these items with the proposed model has been observed by different authors, Lambert et al. (2004), Lo Coco et al. (2008), Vacarezza, Florenzano, & Trapp, 2008, amongst others. Items 11 and 26, however, are used as indicators of the use of alcohol and drugs. Thus the abovementioned authors suggest these items be retained for a qualitative evaluation of these characteristics. As for item 14 in respect of work, the hypothesis is that the results are related to the adaptation of the item *I work/study too much*, which could be modified to *I work/*

study excessively, considering that, for the majority of the Brazilian population, work is highly valued and is not, culturally, regarded as a trigger of symptom distress.

As for the reliability evidence of the OQ-45, the results obtained are consistent with the instrument's original proposal (Lambert et al., 1996) which obtained indices of internal consistency, Cronbach's alpha, ranging from .70 to .93 for specific factors and .92 for the total scale. It should be stressed that similar indices were achieved by different researchers evaluating versions adapted for other countries (Chiappelli, Lo Coco, Gullo, Bensi, & Prestano, 2008; Wennberg, Philips, & de Jong, 2010). However, it is understood that the objectives established in the present study were satisfactorily attained. It may be concluded that the tool presented good psychometric properties for evaluating the Brazilian population, demonstrating its promise for professional use.

Further research is required to establish norms of interpretation for the scale's raw scores, thereby ascribing significance to the results obtained when applying the OQ-45.2 In this sense, it is recommended that the IRT be employed using the Items-Person Map procedure (Peixoto & Nakano, 2014), as well as the performance of further studies seeking further validity evidence for the OQ-45 for the Brazilian population (AERA, APA, & NCME, 2014). Lastly, it is important to continue the evaluation of the instrument's performance using clinical samples that are regionally more diversified.

Abishe, D. R. (2008). *Cross-cultural comparison of the Outcome Questionnaire*. Tese de doutorado em filosofia. Universidade de Oklahoma. Norman: Oklahoma. ProQuest: UMI 3304221.

References

- AE RA – American Educational Research Association, APA – American Psychological Association, NCME – National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Antony, M. M. & Barlow, D. H. (2010). *Handbook of assessment and treatment planning for psychological disorders* (2^a ed.). New York: Guilford Press.
- Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch Model*. Mahwah, NJ: LEA.
- Boswell, D. L., White, J. K., Sims, W. D., Harrist, R. S., & Romans, J. S. (2013). Reliability and validity of the Outcome Questionnaire-45.2. *Psychological Reports: Mental & Physical Health*, 112(3), 689-693. <http://dx.doi.org/10.2466/02.08.PR0.112.3.689-693>
- Brum, E. H. M., Frizzo, G. B., Gomes, Grill. A., Silva, M. R., Sousa, D. D., & Piccinini, C. A. (2012). Evolução dos modelos de pesquisa em psicoterapia. *Estudos de Psicologia*, 9(2), 259-269. <http://dx.doi.org/10.1590/S0103-166X2012000200012>
- Butcher, J., Derksen, J., Sloore, H., & Sirigatti, S. (2003). Objective personality assessment of people in diverse cultures: European adaptations of the MMPI-2. *Behaviour Research and Therapy*, 41(7), 819-840. [http://dx.doi.org/10.1016/S0005-7967\(02\)00186-9](http://dx.doi.org/10.1016/S0005-7967(02)00186-9)
- Carvalho, L. F. & Rocha, G. M.A. (2009). Tradução e adaptação cultural do Outcome Questionnaire (OQ-45). *Psico-USF*, 14(3), 309-316. <http://dx.doi.org/10.1590/S1413-82712009000300007>

- Chiappelli, M., Lo Coco, G., Gullo, S., Bensi, L., & Prestano, C. (2008). L'Outcome-Questionnaire 45.2. The Outcome Questionnaire 45.2. Italian validation of an instrument for the assessment of psychological treatments. *Epidemiologia E Psichiatria Sociale*, *17*, 152-216.
- Cook K. F., Kallen, M. A., & Amtmann D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Qual Life Res*, *18*(4), 447-460. <http://dx.doi.org/10.1007/s11136-009-9464-4>
- Del Prette, G. (2011). *Objetivos analítico-comportamentais e estratégias de intervenção nas interações com a criança em sessões de duas renomadas terapeutas infantis*. Tese de doutorado não publicada. Universidade de São Paulo.
- Doerfler, L. A., Addis, M. E., & Moran, P. W. (2002). Evaluating mental health outcomes in an inpatient setting: Convergent and divergent validity of the OQ-45 and BASIS-32. *The Journal of Behavioral Health Services & Research*, *29*(4), 394-403. <http://dx.doi.org/10.1007/BF02287346>
- Harmon, C., Hawkins, E. J., Lambert, M. J., Slade, K., & Whipple, J.S. (2005). Improving outcomes for poorly responding clients: The use of Clinical Support Tools and feedback to clients. *A Journal of Clinical Psychology*, *61*(2), 175-185. <http://dx.doi.org/10.1002/jclp.20109>
- Hatfield, D. & Ogles, B. M. (2004). The Use of Outcome Measures by Psychologists in Clinical Practice. *Professional Psychology: Research & Practice*, *35*(5), 485-491. <http://dx.doi.org/10.1037/0735-7028.35.5.485>
- Honda, G. C., Peixoto, E. M., Rocha, G. M., & Enéas, M. L. E. (2015). Psicoterapia breve psicodinâmica no processo de formação profissional. In T. V. Santeiro & G. M. A. Rocha (Orgs.). *Clínica de orientação psicanalítica: compromissos, sonhos e inspirações no processo de formação* (pp. 76-80). São Paulo: Vetor.
- Huag, S., Puschner, B., Lambert, M. J., & Kordy, H. (2004). Veraderungs messung in der psychotherapie mit dem Ergebnis Frage Bogen (EB-45) Assessment of change in psychotherapy with the German version of the Outcome Questionnaire (OQ-45). *Zeitschrift für Differentielle und Diagnostische Psychologie*, *25*, 141-151.
- Jong, K., Nugter, M. A., Polak, M. G., Wagenborg, J. E. A., Spinhoven, P., & Heiser, W. J. (2007). The Outcome Questionnaire (OQ-45) in a Dutch population: A cross-cultural validation. *Clinical Psychology and Psychotherapy*, *14*(4), 288-301. <http://dx.doi.org/10.1002/cpp.529>
- Lambert, M. J. (1983). Introduction to assessment of psychotherapy outcome: Historical perspective and current issues. In: M. J. Lambert, E. R. Christensen, & S. S. DeJulio (Eds.), *The assessment of psychotherapy outcome* (pp. 3-32). New York: John Wiley and Sons.
- Lambert, M. J., Burlingame, G. N., Umphress, V., Hanse, N.B., Vermeersh, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy*, *3*, 249-258. [http://dx.doi.org/10.1002/\(SICI\)1099-0879\(199612\)3:4<249::AID-CPP106>3.0.CO;2-S](http://dx.doi.org/10.1002/(SICI)1099-0879(199612)3:4<249::AID-CPP106>3.0.CO;2-S)
- Lambert, M. J. & Finch, A. E. (1999). The Outcome Questionnaire. M. E. Maruish (Ed.). *The use of psychological testing for treatment planning and outcomes assessment* (pp. 831-869). New Jersey: Lawrence Erlbaum Associates, Publishers.
- Lambert, M.J., Hansen, N.B., & Finch, A.E. (2001). Patient-focused research: Using patient outcome data to enhance treatment effects. *Journal of Consulting and Clinical Psychology*, *69*, 159-172. <http://dx.doi.org/10.1037/0022-006X.69.2.159>
- Lambert, M. J. (1994). Assessing psychotherapy outcomes and processes. In A. E. Bergin and S. L. Garfield (Eds.). *Handbook of psychotherapy and behavior change*, *4*, 72-113. New York: John Wiley and Sons.
- Lambert, M. J., Morton, J. J., Hartfield, D., Harmon, C., Hamilton, S. Reid, R. C., Shimokawa, K., Christopherson, C., & Burlingame, G. M. (2004). *Administration and Scoring Manual for the OQ-45.2 Outcome Questionnaire*. Salt Lake City: American Professional Credentialing Services.
- Lambert, J. M., Smart, D. W., Campbell, M. P., Hawkins, E. J., Harmon, C., & Slade, K. L. (2006). Psychotherapy outcome, as measured by the OQ-45, in African American, Asia/Pacific Islander, Latino/a, and Native American clients compared with matched caucasian clients. *Journal of College Student Psychotherapy*, *20*(4), 17-29. http://dx.doi.org/10.1300/J035v20n04_03
- Lara, D. & Alexis, S. (2014). ¿Matrices Policóricas/Tetracóricas o Matrices Pearson? Un estudio metodológico. *Revista Argentina de Ciencias del Comportamiento*, *6*(1), 39-48.
- Lara, C., Cruz, C., Vacarezza, A., Florenzano U. R., & Trapp, A. (2008). Análisis comparativo de dos instrumentos de evaluación clínica: OQ45 e InterRAI – Salud Mental. *Revista Chilena de NeuroPsiquiatria*, *46*(3), 192-198. <http://dx.doi.org/10.4067/S0717-92272008000300004>
- Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.
- Linacre J. M. (2015). A user's guide to Winsteps Ministep: Rasch-model computer programs. Retrieved from: <http://www.winsteps.com>
- Lo Coco, G. L., Chiappelli, M., Bensi, L., Gullo, S., Prestano, C., & Lambert, M. J. (2008). The factorial structure of the Outcome Questionnaire-45: A study with an Italian sample. *Clinical Psychology and Psychotherapy*, *15*, 418-423. <http://dx.doi.org/10.1002/cpp.601>
- Machado, P. P. P. & Fassnacht, D. B. (2015). The Portuguese version of the Outcome Questionnaire (OQ-45): Normative data, reliability, and clinical significance cut-offs scores. *Psychology and Psychotherapy: Theory, Research and*

- Practice* (2015), 88, 427-437. <http://dx.doi.org/10.1111/papt.12048>
- Milfont, T. L. & Fisher, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3(1), 111-121.
- Mueller, R. M., Lambert, M. J., & Burlingame, G. M. (1998). Construct validity of the Outcome Questionnaire: A confirmatory factor analysis. *Journal of Personality Assessment*, 70(2), 248-262. http://dx.doi.org/10.1207/s15327752jpa7002_5
- Muthén, L. K. & Muthén, B. O. (2012). *Mplus User's Guide* (7th ed.). Los Angeles, CA.
- Parra, G. & Bergen, V. (2002). OQ 45-2 Cuestionario para evaluación de resultados y evolución en psicoterapia: adaptación, validación, e indicaciones para su aplicación e interpretación. *Revista de Terapia*, 20, 161-176.
- Pasquali, L. (2010). Testes Referentes a Construto: Teorias e Modelos de Construção. In: L. Pasquali et al. (Orgs.), *Instrumentação Psicológica: Fundamentos e prática* (pp. 165-198). Porto Alegre: Artmed.
- Peuker, A. C., Habigzang, L. F., Koller, S. H., & Araujo, L. B. (2009). Avaliação de processo e resultado em psicoterapias: uma revisão. *Psicologia em Estudo*, 14(3), 439-445. <http://dx.doi.org/10.1590/S1413-73722009000300004>
- Pieta, M. A. M., Siegmund, G., Gomes, W. B., & Gauer, G. (2015). Desenvolvimento de protocolos para acompanhamento de psicoterapia pela Internet. *Contextos Clínicos*, 8(2), 128-140. <http://dx.doi.org/10.4013/ctc.2015.82.02>
- Peixoto, E. M. & Nakano, T. C. (2014). Problemas e perspectivas na utilização dos testes psicológicos em psicologia do esporte. In: C. R. Campos, T. C. Nakano. (Org.). *Avaliação Psicológica direcionada a populações específicas: técnicas, métodos e estratégias* (pp. 201-23). São Paulo: Vetor.
- Primi, R., Carvalho, L. F., Miguel, F. K., & Muniz, M. (2010). Resultado dos fatores da BFP por meio da Teoria de Resposta ao Item: interpretação referenciada no item. In: C. H. S. S., Nunes, C. S., Hutz, & M. F. O., Nunes. (Org.). *Bateria Fatorial de Personalidade (BFP): manual técnico* (pp. 153-170). São Paulo: Casa do Psicólogo.
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696. <http://dx.doi.org/10.1080/00273171.2012.715555>
- Rios, J. & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, 26(1), 08-116. [10.7334/psicothema2013.260](https://doi.org/10.7334/psicothema2013.260)
- Rodríguez, M. A. von B. (2000). *Investigación Empírica em Psicoterapia: Validación del Cuestionario de Resultados Terapéuticos OQ-45.2*. Tesis para optar al grado de Licenciado em Psicologia. Universidad Nacional Andres Bello. Santiago do Chile.
- Santos, D. & Primi, R. (2014). *Social and emotional development and school learning: a measurement proposal in support of public policy*. Technical report for Organization for Economic Cooperation and Development (OCDE) Rio de Janeiro State Education Department (SEEDUC) and Ayrton Senna Institute. São Paulo: Ayrton Senna Institute.
- Serralta, F. B., Nunes, M. L. T., & Eizirik, C. L. (2007). Elaboração da versão em português do Psychotherapy Process Q-Set. *Revista Brasileira de Psiquiatria*, 29(1), 44-55. <http://dx.doi.org/10.1590/S0101-81082007000100011>
- Silva, S. M. (2013). *Escala de Avaliação de Resultados (Outcome Questionnaire) – OQ-45.2: Validade e precisão*. Tese de Doutorado [não publicada]. Universidade de São Paulo.
- Umpress, V. J., Lambert, M. J., Smart, D. W., Barlow, S. H., & Clouse, G. (1997). Concurrent and construct validity of the Outcome Questionnaire. *Journal of Psychoeducational Assessment*, 15, 40-55. <http://dx.doi.org/10.1177/073428299701500104>
- Wennberg, P., Philips, B., & de Jong, K. (2010). The Swedish version of the Outcome Questionnaire (OQ-45): Reliability and factor structure in a substance abuse sample. *Psychology and Psychotherapy: Theory, Research and Practice*, 83, 325-329. <http://dx.doi.org/10.1348/147608309X478715>

Authors:

Sonia Maria da Silva – Academician, Universidade Guarulhos.
 Irai Cristina Boccato Alves – PhD, Instituto de Psicologia da USP.
 Evandro Morais Peixoto – PhD, Pontifícia Universidade Católica de Campinas.
 Glaucia Mitsuko Ataka Rocha – PhD, Instituto de Psicologia da USP.
 Tatiana de Cássia Nakano – PhD, Pontifícia Universidade Católica de Campinas.

Correspondence:

Evandro Morais Peixoto
 Street Liliane Regina, 3
 06386-300 Carapicuíba, SP, Brazil
[<epeixoto_6@hotmail.com>](mailto:epeixoto_6@hotmail.com)

Received: 2016, July 15

Accepted: 2016, Sept. 22