ISSN 1697-2600 print
ISSN 2174-0852 online
2011, Vol. 11, Nº 3, pp. 473-493

The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis[1]

Julio Sánchez-Meca[2], José A. López-Pina, José A. López-López,
Fulgencio Marín-Martínez, Ana I. Rosa-Alcázar, and Antonia Gómez-Conesa
(*Universidad de Murcia, Spain)*

ABSTRACT. The Maudsley Obsessive-Compulsive Inventory (MOCI) is one of the most used tests in clinical psychology for assessing the obsessive and compulsive symptoms in psychiatric patients and as a screening tool in nonclinical population. A reliability generalization meta-analysis was carried out with the purpose of studying how test scores reliability varies along different applications of this test. An exhaustive search of the literature enabled us to select 51 studies that reported some reliability estimate and, by means of the *KR*-21 formula, we were able to increase the database to 308 internal consistency reliability estimates. On average, the internal consistency reliability of test scores was .76 for the original 30-items version. The reliability coefficients exhibited a large heterogeneity. The analyses of moderator variables revealed a predictive model composed of two predictors: the standard deviation and the mean of test scores. Our results confirmed that reliability is not a property of the test itself, but it varies from one application to the next. As a consequence, the erroneous practice of inducing reliability from previous studies should be avoided in psychological research.

KEYWORDS. MOCI. Maudsley Obsessive-Compulsive Inventory. Reliability. Coefficient alpha. Meta-analysis.

RESUMEN. El Inventario de Obsesiones y Compulsiones de Maudsley (MOCI) es uno de los tests más utilizados en psicología clínica para evaluar las obsesiones y compulsiones en pacientes psiquiátricos y como un instrumento de cribado en pobla-

[1] This article was supported by a grant of the Fondo de Investigación Sanitaria (FIS), Instituto de Salud Carlos III of the Spanish Government and by FEDER funds (Project nº PI07/90384).
[2] Correspondence: Dpto. Psicología Básica y Metodología. Facultad de Psicología. Campus de Espinardo. 30100 Murcia (Spain). E-mail: jsmeca@um.es. http://www.um.es/metaanalysis.

ción no clínica. Con objeto de estudiar cómo varía la fiabilidad de las puntuaciones del test, se ha realizado un estudio de generalización de la fiabilidad. Una búsqueda exhaustiva de la literatura nos permitió seleccionar 51 estudios que reportaron alguna estimación de la fiabilidad y, mediante la fórmula *KR*-21, pudimos incrementar la base de datos hasta 308 estimaciones de la fiabilidad por consistencia interna. En promedio, la fiabilidad por consistencia interna de las puntuaciones fue de 0,76 para la versión original de 30 ítems. Los coeficientes de fiabilidad exhibieron una gran variabilidad. Los análisis de variables moderadoras revelaron un modelo predictivo con dos predictores: la desviación típica y la media de las puntuaciones del test. Nuestros resultados confirman que la fiabilidad no es una propiedad inherente al test, sino que varía de una aplicación a otra del mismo. En consecuencia, la práctica errónea de inducir la fiabilidad a partir de estudios previos debe ser erradicada de la investigación psicológica.

PALABRAS CLAVE. MOCI. Inventario de Obsesiones y Compulsiones de Maudsley. Fiabilidad. Coeficiente alfa. Meta-análisis.

The obsessive-compulsive disorder (OCD) is one of the most common anxiety disorders, with a prevalence rate around 1-4% (Rosa-Alcázar, Sánchez-Meca, Gómez-Conesa, and Marín-Martínez, 2009). One of the psychological tests most commonly used for assessing the obsessive and compulsive symptoms is the Maudsley Obsessive-Compulsive Inventory (MOCI) developed by Hodgson and Rachman (1977). The MOCI is a self-report questionnaire with true-false format developed for evaluating the type of obsessive-compulsive symptoms and discriminating obsessive patients from other neurotic patients and from nonclinical people. The test is composed of 30 dichotomous items, so that the total score for a subject will range between 0 (*absence of symptoms*) and 30 (*maximum presence of symptoms*). The original version has four subscales: *Checking* (9 items), *Cleaning* (11 items), *Slowness* (7 items), and *Doubting* (7 items). Note that the sum of the items for the four subscales is 34, not 30, because four items are included in two subscales (Hodgson and Rachman, 1977).

The MOCI can be applied to adults as well as children and adolescents. In addition, it has been applied for assessing obsessive and compulsive symptoms not only in patients with OCD, but also for other patient populations (*e.g.*, depressive patients), and as a screening tool for nonclinical populations (Einstein and Menzies, 2006). The MOCI is also a very sensitive instrument to therapeutic change and, as a consequence, it has been applied very frequently in empirical studies evaluating the effectiveness of psychological and/or pharmacological treatments for patients with OCD (Thordarson *et al.*, 2004).

Since it was developed, the MOCI has been adapted to many different languages and cultures. So, there are versions of the MOCI adapted to the Spanish (Ávila-Espada, 1986; Fonseca-Pedrero, Paíno-Piñeiro, and Lemos-Giraldez, 2007; Guilera, Gómez-Benito, Tomás, and Carrera, 2005), German (Müller *et al.*, 1997), French (Hantouche and Guelfi, 1993), Dutch (Meesters, 1997), Italian (Sanavio and Vidotto, 1985), Norwegian (Støylen, Larsen, and Kvale, 2000), Icelandic (Smári, Bjarnason, Porleifsson, Sturludóttir, and Hafsteinsdóttir 1994), Turkish (Erol and Savasir, 1989), Hebrew (Zohar, LaBuda, and

Moschel-Ravid, 1995), Korean (Min and Won, 1999), Chinese (Chan, 1990), and Japanese (Tadai, Nakamura, Okazaki, and Nakajima, 1995).

To be useful, a psychological test must show good psychometric properties in terms of reliability and validity. This research is focused on score reliability for the MOCI. From the classical test theory, reliability can be defined as the consistency of measurement over the conditions of testing (Anastasi and Urbina, 1997). Such conditions include time sampling (*e.g.,* test-retest correlation), content sampling (*e.g.,* coefficient alpha), and scorer/rater differences (*e.g.,* intraclass correlation). Several psychometric studies have analyzed the score reliability of the MOCI. In the original study, Hodgson and Rachman (1977) applied the MOCI to 50 OCD and 50 neurotic patients and obtained coefficients alpha of .70, .80, .70, and .70 for the *Checking*, *Cleaning*, *Slowness*, and *Doubting* subscales, respectively. Regarding test-retest reliability, they obtained a value of .80 on the sample of 50 OCD patients. Later, Rachman, and Hodgson (1980) examined the psychometric properties of the MOCI on a sample of 50 neurotic patients and a sample of 50 nonclinical people, finding similar reliability estimates, but verifying the suspect that the *Slowness* subscale did not have good properties. Other psychometric studies have found similar results to the ones obtained by Hodgson and Rachman (1977), such as those of Sternberger and Burns (1990), Emmelkamp, Kraaijkamp, and Van den Hout (1999), Støylen *et al.* (2000), and Li and Chen (2007). On the other hand, other studies have suggested deleting the *Slowness* subscale and maintaining the other three (*e.g.,* Chan, 1990; Sanavio and Vidotto, 1985).

In Spain, the first adaptations were carried out by Ávila-Espada (1986) and Raich (1996), but the first psychometric study with Spanish population was published by Guilera *et al.* (2005). They applied a Spanish version of the scale to a sample of 312 university students, obtaining a coefficient alpha for the total scale of .81. Guilera *et al.* (2005) did not report coefficients alpha for the subscales, but in a recent study Fonseca-Pedrero, Lemos-Giráldez, Paíno-Piñeiro, Villazón-García, and Muñiz (2010) applied the MOCI to a sample of 508 nonclinical children and adolescents, obtaining coefficients alpha of .67, .42, .45, and .50 for *Checking*, *Cleaning*, *Slowness*, and *Doubting* subscales, respectively, and an alpha of .75 for the total scale.

The reliability generalization approach

The psychometric theory states that reliability is not a property of the test, but of the test scores obtained in an application of it to a particular sample of participants (Crocker and Algina, 1986; Pedhazur and Schmelkin, 1991). This is because reliability of test scores can change depending on the composition and characteristics of the samples of participants and also on the application context. As reliability varies in each test administration, researchers should report the reliability obtained for the data at hand. However, it is very common that researchers cite reliability estimates obtained in previous studies or normative samples rather than estimating score reliability with the own data. This phenomenon, known as *reliability induction* (Vacha-Haase, Kogan, and Thompson, 2000), does not take into account that the reliability of test scores obtained in a particular sample is only comparable to that obtained in another sample if both have the same composition and variability. And this correspondence is almost never checked

(Crocker and Algina, 1986). Unfortunately, many researchers make the mistake of inducing score reliability from that obtained in previous studies. Some studies have analyzed the reporting practices of score reliability and have found large rates of reliability induction, even reaching percentages of 88.8%, for the Buss Durkee Hostiliy Inventory (Vassar and Hale, 2009), or 94%, for the Spielberger Strait-Trait Anxiety Inventory (Barnes, Harp, and Jung, 1998).

Although there are many studies that induce score reliability, with the studies that report reliability estimates from the data at hand, it is possible to conduct a meta-analytic *reliability generalization* (RG) study. Vacha-Haase (1998) coined the term «reliability generalization» to refer to this type of research. The purpose in an RG study is to obtain an average estimate of the score reliability of the test, to determine whether the reliability coefficients are heterogeneous among themselves and, where appropriate, to examine which characteristics of the test, of the studies and of the participants can account for that heterogeneity (Henson and Thompson, 2002; Rodriguez and Maeda, 2006; Sánchez-Meca and López-Pina, 2008). From the Vacha-Haase's (1998) seminal work, more than 80 RG studies about different psychological tests have been published. For example, RG studies have been carried out about the Beck Depression Inventory (Yin and Fan, 2000), the Eysenck Personality Inventory (Caruso, Witkievitz, Belcourt-Dittloff, and Gottlieb, 2001), the Maslach Burnout Inventory (Aguayo, Vargas, de la Fuente, and Lozano, 2011), and the Hamilton Rating Scale for Depression (López-Pina, Sánchez-Meca, and Rosa-Alcázar, 2009).

Objectives of the study

This overview of the MOCI makes clear that it is a very widely used test in clinical populations and for screening purposes. In addition, the test has been adapted to different languages and cultures. It seems reasonable, then, to expect a large variability in the reliability estimates along different administrations of the MOCI. Therefore, the purpose of this research was to carry out an RG meta-analysis with the aim of obtaining an estimate of the mean reliability of the test scores and to search for characteristics of the studies and of the participants that are affecting the reliability estimates (Botella and Gambara, 2006; Montero and León, 2007). In particular, it was expected that characteristics such as the mean and the standard deviation of the test scores, the age and the target population of the participants (clinical *vs.* nonclinical), and the version of the test (original *vs*. adapted), would affect the score reliability. As internal consistency is the most usual type of reliability reported in the studies, we focused on coefficient alpha as the reliability coefficient. We were also interested in estimating the mean score reliability of the different subscales of the MOCI and in examining whether they have, in general, a good score reliability taking .7 as the criterion (Nunnally and Bernstein, 1994).

One of the most serious problems when carrying out RG studies is the failure of researchers to report reliability estimates from their own data. When the measurement instrument is composed of a set of dichotomous items, however, it is possible to obtain an estimate of the coefficient alpha by the *KR*-21 formula if the study reports the number of items in the test, the total score mean, and the total score standard deviation (Crocker and Algina, 1986).

When the items vary in difficulty, *KR*-21 will underestimate the true coefficient alpha, but the underestimation will be less than 0.03 for coefficients alpha greater than .60 (Hopkins, 1998). Several authors have proposed increasing the number of coefficients alpha in RG studies by estimating them through *KR*-21 when the study fails to report reliability (Henson, Kogan, and Vacha-Haase, 2001; Kieffer and Reese, 2002; Lane, White, and Henson, 2002). At the same time, studies that report coefficient alpha and descriptive statistics for the total test score will make it possible to examine the magnitude of the underestimation produced by *KR*-21, as well as the relationship between *KR*-21 and coefficient alpha. Thus, in their *RG* study on the Coopersmith Self-Esteem Inventory, Lane *et al.* (2002) found a mean underestimation of *KR*-21 equal to .02 and a correlation of .90 between *KR*-21 and coefficient alpha.

Method

Locating studies

An RG study requires selecting the studies that have applied the test and that report a reliability estimate from the data at hand. To be included, studies were required to meet four criteria: a) studies which had applied the MOCI to a sample of participants, b) they had to report a coefficient alpha computed from their own sample data, c) they had to be carried out or published between 1977 and 2008, and d) due to idiom limitations, they had to be written in English, French, Spanish or Portuguese.

To locate the studies, the following electronic databases were consulted: PsycInfo, MedLine, ERIC, and the Scholar Google search engine, by combining «Maudsley AND Obsessiv*» as keywords. In addition, references from 26 meta-analytic reviews about OCD, specialized journals, books and monographs, as well as references from the already retrieved studies, were also consulted. Finally, expert researchers in this topic were contacted in order to locate unpublished studies (or hardly recoverable ones) which could have applied the MOCI.

This search process entailed consulting more than 3,500 references, which allowed us to identify 327 studies which had applied the MOCI. Only 51 studies (15.6%) reported a reliability estimate from the data at hand and, as a consequence, did not induce reliability from previous studies. The remaining 276 studies (84.4%) induced reliability. Two kinds of reliability induction can be distinguished (Shields and Caruso, 2004): reliability induction by omission consists in not making any reference to the test scores reliability, whereas reliability induction by report occurs when the study reports some reliability estimate from previous studies. In our RG meta-analysis, the number of studies that induced reliability by omission and by report was 201 (61.5%) and 75 (22.9%), respectively. The 51 studies that did not induce reliability were the target population in our RG study. With the exception of two studies that were written in Spanish, the 49 remaining studies were written in English. The continent most represented in this database was Europe with 28 studies (54.9%), followed by North America with 14 studies (27.4%), Asia with 5 studies (9.8%), and Oceania with 4 Australian studies (7.8%).

Coding study characteristics

In order to explore how study characteristics can affect the measurement error when the MOCI is applied, the following moderator variables were coded in the studies: a) standard deviation of the test scores (for total scale and subscales); b) mean of the test scores (for total scale and subscales); c) test version (original *vs*. other); d) test length (30 items, 37 items, other); e) mean age of participants (in years); f) standard deviation of the age of participants (in years); g) gender distribution in the sample (% male); h) sample ethnicity (% Caucasian); i) target population of the sample (nonclinical, University students, clinical, and subclinical); j) disorder of the participants (OCD, eating disorders, and other); k) mean of disorder history (in years); l) standard deviation of disorder history (in years); m) study focus (psychometric vs. substantive); n) focus of the psychometric studies (the MOCI *vs*. other tests); o) publication year of the study; p) researcher affiliation (psychology *vs.* psychiatry), and q) sample size. Together with these moderator variables, coefficients alpha and test-retest correlations were obtained for the total scale and for the subscales when they were reported in the studies.

The reliability of the coding process of the study characteristics was checked by selecting a random sample of 20% of the studies that had applied the MOCI. This sample of studies was double-coded by two independent coding teams. In general, the intercoder agreement was satisfactory, with kappa coefficients ranging between .73 and .98 for the qualitative characteristics, and intraclass correlations ranging between .75 and .96 for the continuous variables.

Statistical analyses

Separate meta-analyses were carried out for coefficients alpha and test-retest correlations, as they are two different types of reliability. In order to increase the database of internal consistency coefficients, the *KR*-21 formula was applied to those studies that did not report a coefficient alpha but reported the number of items (J) and the mean ($\overline{X}$) and standard deviation (S_x) of the test scores for the total scale. Thus, it was possible to estimate coefficient alpha by applying the formula:

$$KR-21=\frac{J}{J-1}\left[1-\frac{\overline{X}\left(J-\overline{X}\right)}{JS_x^2}\right] . \quad (1)$$

The statistical analyses with coefficients alpha and *KR*-21 estimates were carried out after transforming these coefficients by means of the Bonett's (2002) formula: $T = Ln(1-\hat{\alpha}) Ln$ being the natural logarithm, T being the transformed coefficient and $\hat{\alpha}$ being the coefficient alpha or its estimation by *KR*-21 formula. To facilitate interpretation of results, the average T values and their confidence limits were back-transformed to present the results in the original metric of reliability coefficients, by means of the transformation formula: $\hat{\alpha}= 1 - e^T$. Test-retest correlations were transformed by Fisher's Z. Using a transformation formula of the reliability coefficients allows to normalize its distribution and to stabilize their variances (Botella, Suero, and Gambara, 2010; Rodriguez and Maeda, 2006). To obtain summary statistics of reliability coefficients, a random-

effects model was assumed and, as a consequence, the reliability coefficients were weighted by the inverse variance. The heterogeneity exhibited by the reliability estimates was assessed by the Q test and the I^2 index. Finally, the influence of moderator variables was examined by applying regression analyses for the continuous variables and analyses of variance (ANOVAs) for the qualitative ones from a mixed-effects model (Borenstein, Hedges, Higgins, and Rothstein, 2009; Sánchez-Meca and Marín-Martínez, 2008).

Results

First, summary statistics for reliability estimates (coefficients alpha and test-retest correlations) will be presented. Next, an analysis of the equivalence of *KR*-21 formula as an estimator of coefficient alpha is presented. This section will finish with a description of the analyses of moderator variables and the proposal of a predictive model.

Estimating the mean reliability

A total of 51 studies reported a reliability estimate with the data at hand. Out of the 51 studies, 39 of them reported a coefficient alpha for the total scale in its original 30-items version, whereas 5 studies reported coefficients alpha from the 37-items (Turkish) version. In addition, five studies reported a test-retest correlation, three of them for the 30-items version and the other two for the 37-items version. The results are presented separately for both test lengths. Table 1 shows the main summary statistics for coefficients alpha and test-retest reliability coefficients. The 30-items version presented a (weighted) mean coefficient alpha of .76 (confidence limits: .75 and .78), clearly over the cut off value of .70 to be considered an acceptable reliability (Nunnally and Bernstein, 1994). The Turkish version, with 37 items, was even better, reaching an average coefficient alpha of .82, although based on five studies only. Table 1 also shows the results for the three Spanish studies that reported a coefficient alpha, with a mean of .77 (confidence limits: .75 and .79).

TABLE 1. Summary results for the coefficients alpha, test-retest correlations, and *KR*-21 estimates obtained from the studies.

Test version / subscale	*k*	*Minimum*	*Maximum*	*Unweighted mean*	*Weighted mean*	*95% C. I.* *Lower*	*Upper*	*Q*	I^2
Total scale (30-items version):	39	.61	.87	.76	.76	.75	.78	197.02**	80.7
Checking subscale	22	.50	.84	.64	.64	.61	.67	73.65**	71.5
Cleaning subscale	23	.39	.87	.55	.56	.51	.60	110.46**	80.1
Slowness subscale	16	-.12	.70	.36	.40	.34	.47	55.54**	73.0
Doubting subscale	19	.28	.86	.56	.57	.51	.62	135.01**	86.7
Total scale (37-items version):[a]	5	.80	.86	.82	.82	.80	.83	5.77	30.7
Total scale (Spanish adaptation):[a]	3	.75	.81	.77	.77	.75	.79	7.03*	71.6
Total scale test-retest reliability:[a, b]									
30-items version	3	.69	.80	.73	.70	.65	.75	2.44	18.2
37-items version	2	.81	.88	.84	.85	.81	.88	3.64	72.5
Coefficients alpha + *KR*-21 estimates:									
30-items version	289	.002	.97	.68	.74	.72	.75	1956.55**	85.3
37-items version	19	.49	.89	.73	.77	.73	.80	82.54**	78.2

Note. With the exception of the rows for test-retest correlations, all of the remaining results refer to internal consistency coefficients (coefficients alpha and *KR*-21 estimates).
[a] A fixed-effects model was applied because of the small number of coefficients.
[b] The test-retest coefficients were previously transformed into the Fisher's Z and then they were back-transformed to the correlation metric.
k: number of coefficients. * $p < .05$. ** $p < .001$. *Q*: heterogeneity statistic. I^2: heterogeneity index.

Some studies reported coefficients alpha separately for the four subscales of the MOCI, obtaining mean reliability estimates of .64 for the *Checking* subscale, .56 for the *Cleaning* subscale, .40 for the *Slowness* subscale, and .57 for the *Doubting* subscale. The four subscales exhibited a mean internal consistency clearly under the cut off of .70. Further, their confidence limits did not include that cut off value. Out of the four subscales, and in agreement with previous results, the *Slowness* subscale exhibited the poorest internal consistency.

With regards to test-retest reliability, the three studies that reported reliability estimates for the 30-items original version achieved a mean reliability of .70, whereas the two Turkish studies (37-items version) obtained a larger mean reliability of .85.

KR-21 estimates

As the MOCI is composed of dichotomous items, the *KR*-21 formula was applied to estimate the coefficient alpha when the study failed to report an internal consistency estimate but reported the number of items of the scale and the mean and the standard deviation of test scores. In order to examine the adequacy of using *KR*-21 formula as an estimate of coefficient alpha, we selected the 34 studies that reported a coefficient alpha with the data at hand in addition to the number of items, the mean and the standard deviation of test scores. With equation 1) we calculated the *KR*-21 estimate, obtaining a mean of .74. Comparing this mean with that obtained with the 34 coefficients alpha, .76, the underestimation of *KR*-21 formula was of .02 only, a very similar difference to that obtained by Henson and Thompson´s (2002) RG study, and also in agreement with the maximum .03 underestimation stated by Hopkins (1998). Although the underestimation can be considered negligible, we found statistically significant differences between the mean coefficient alpha and the mean *KR*-21 [T (33) = 3.29, p = .002]. However, the Pearson correlation coefficient between coefficients alpha and *KR*-21 estimates was of large magnitude and statistically significant (r = .68, p < .001). Figure 1 presents a scatter plot of the relationship between coefficients alpha and *KR*-21 estimates. Thus, *KR*-21 estimates and coefficients alpha share an important amount of variability and, as a consequence, *KR*-21 is a useful formula to estimate coefficient alpha in *RG* studies when the empirical studies fail to report an internal consistency coefficient with the data at hand.

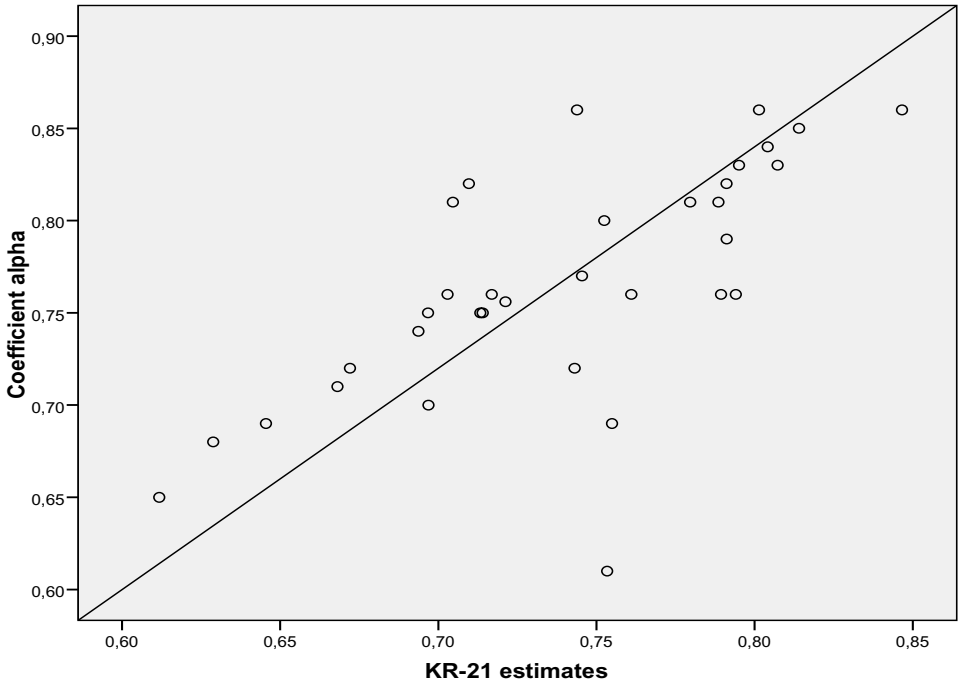

FIGURE 1. Scatter plot of the relationship between coefficients alpha and *KR*-21 estimates obtained with the 34 studies that reported a coefficient alpha estimate and the mean and the standard deviation of the test scores.

In our RG meta-analysis we found 264 studies that failed to report coefficients alpha, but reported information to apply the *KR*-21 formula. Thus, we were able to increase the original database of 44 coefficients alpha to 308 internal consistency estimates, by adding the new 264 *KR*-21 estimates to the original coefficients. As Table 1 shows, the mean internal consistency obtained was very similar to that obtained for the original database of coefficients alpha: .74 for the 30-items version.

Table 1 also presents the results of the *Q* statistics and the I^2 indices for assessing the variability exhibited by reliability estimates. As Table 1 shows, the coefficients alpha showed a statistically significant heterogeneity, with I^2 values over 70% for the total scale and for the subscales. As expected, adding *KR*-21 estimates to the original coefficients alpha led to a large heterogeneity. As a consequence, analyses to explain the coefficients variability were in order. To accomplish this objective taking advantage of the maximum information available, the analyses of moderator variables were carried out taking the complete database resulting of adding the *KR*-21 estimates to the original coefficients alpha.

Searching for moderator variables

To explore the study and sample characteristics related to reliability, we used the 289 coefficients alpha and *KR*-21 estimates obtained from the studies that applied the 30-items version of the MOCI. This composite of reliability estimates was the dependent variable, whereas the moderator variables previously coded were the predictors that could explain the variability exhibited by the reliability coefficients. To examine the influence of continuous moderator variables, simple regression analyses were carried out, whereas for exploring the influence of qualitative moderators, ANOVAs were accomplished. In all cases, mixed-effects models were assumed taking as the dependent variable the Bonett's (2002) transformation applied to the internal consistency estimates.

TABLE 2. Weighted simple regression analyses of the continuous moderator variables on the internal consistency reliability estimates (coefficients alpha + *KR*-21 estimates).

Moderator variable	*k*	*Min.*	*Max.*	*Mean*	*Median*	b_j	*Z*	*p*	Q_E	R^2
SD of test scores	280	1.55	10.33	4.60	4.60	-.35	-33.52	< .001	364.91**	.946
Mean of test scores	280	2.30	28.55	10.30	9.05	-.01	-2.25	.025	1892.03**	.002
Mean age (in years)	267	14.45	63.90	31.23	32.00	.01	3.67	< .001	1651.04**	.028
SD of the age (in years)	220	.51	19.20	8.36	9.19	.01	1.50	.132	1519.18**	.006
Gender (% male)	262	.00	100.00	35.41	38.00	.00	.98	.327	1742.34**	.0
Ethnicity (% Caucasian)	28	.00	100.00	63.04	80.50	-.00	-2.49	.013	211.74**	.024
Mean of disorder history (in years)	79	1.08	26.26	11.62	10.95	.00	.59	.552	390.53**	.0
SD of disorder history (in years)	61	.78	18.35	7.54	7.53	.01	.71	.477	247.02**	.019
Publication year	289	1977	2009	1999	2000	-.00	-.63	.528	1934.55**	.0

Notes. k: number of coefficients. Min. and Max.: minimum and maximum values, respectively. b_j: unstandardized regression coefficient for the moderator variable. Z: statistical test for testing the statistical significance of the moderator variable. Q_E: statistic for testing the model specification.. R^2: proportion of variance accounted for by the moderator variable. ** $p < .001$.

TABLE 3. Weighted ANOVAs of the qualitative moderator variables on internal consistency reliability estimates (coefficients alpha + *KR*-21 estimates).

Moderator variable	*k*	*Weighted mean*	*95% C. I.* *Lower*	*Upper*	Q_{wj}	Q_B	*p*	R^2
Test version:						5.13	.023	.023
Original	162	.75	.73	.76	1082.68**			
Other	127	.72	.70	.74	807.25**			
Study focus:						2.06	.151	.0
Psychometric	51	.75	.73	.78	195.68**			
Substantive	238	.73	.72	.74	1758.00**			
Psychometric focus:						9.12	.002	.022
MOCI	17	.72	.69	.75	90.94**			
Other	34	.77	.75	.79	92.06**			
Target population:						24.67	< .001	.001
Non clinical	65	.68	.65	.71	261.52**			
University students	56	.77	.74	.79	753.39**			
Clinical	155	.75	.73	.76	825.39**			
Subclinical	12	.68	.59	.75	53.36**			
Disorder:						9.20	.010	.099
OCD	98	.73	.70	.76	401.94**			
Eating disorders	29	.80	.76	.83	171.06**			
Other	41	.71	.67	.75	214.29**			
Researcher affiliation:						1.27	.259	.0
Psychology	164	.74	.732	.76	1274.15**			
Psychiatry	96	.73	.709	.75	519.60**			

Notes. *k*: number of coefficients. Q_{wj}: within-category statistic for testing the model specification. Q_B: between-categories statistic for testing the statistical significance of the moderator variable. R^2: proportion of variance accounted for by the moderator variable. ** $p < .00$

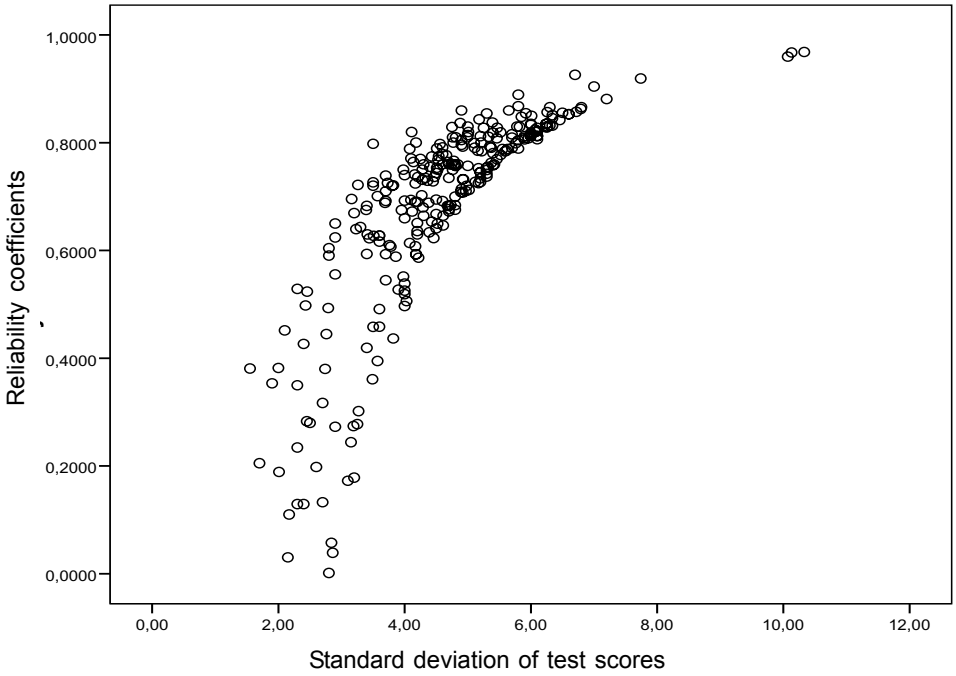

FIGURE 2. Scatter plot of the relationship between reliability coefficients and test scores variability.

Table 2 presents the results of the weighted simple regression analyses for the continuous moderator variables, whereas Table 3 presents those of the ANOVAs applied for the qualitative moderator variables. Note that the sign of the regression coefficients obtained in the regression models have to be interpreted at the inverse, as the Bonett's transformation of the reliability coefficients produces a scale reversion. As psychometric theory predicts, there was a positive, highly significant relationship between the standard deviation of test scores and the reliability estimates ($Z = 33.52$, $p < .001$), with a 94.6% of variance accounted for (see Table 2). The strong positive relationship between reliability and test scores variability can be observed in the scatter plot presented in Figure 2. In addition, a negative, statistically significant relationship was found between the mean of the test scores and the reliability estimates ($p = .025$), although with a negligible percentage of variance accounted for of 0.2%. The mean age of the participants also showed a negative, statistically significant relationship with the reliability coefficients ($p < .001$), but with a small percentage of variance accounted for (2.8%). The sample ethnicity, defined as the percentage of Caucasians, also achieved a negative, significant result ($p = .013$) and a small percentage of variance accounted for (2.4%). The remaining continuous moderator variables here tested did not reach the statistical significance.

With regards to the qualitative moderator variables, Table 3 shows the results of the weighted ANOVAs applied on the reliability coefficients. Several moderator variables

reached the statistical significance, but with very small proportions of variance accounted for (all of them under .10). This was the case for the test version (p = .023), which showed a better mean reliability coefficient for the original test than for adaptations of the test to other languages and/or cultures. The target population also reached a statistically significant result (p < .001), with better reliability estimates when participants were University students or psychiatric patients, and with lower reliability values for nonclinical or subclinical participants. Moreover, when the studies composed of psychiatric patients were classified as a function of the disorder type, there were also found significant differences between OCD patients (mean reliability = .73), patients with eating disorders (mean = .80), and other psychiatric disorders (mean = .71). Some of the studies were focused on examining the psychometric properties of different tests. When these studies were classified into studies about the MOCI versus other tests, a statistically significant difference was found between the two mean reliability coefficients, with a higher mean for studies focused on other tests (mean = .77) than that for those focused on the MOCI (mean = .72). Finally, as Table 3 shows, there were no statistically significant differences when the studies were classified as a function of their focus (psychometric versus substantive) or by the main researcher affiliation (psychology versus psychiatry).

Although some of the moderator variables here analyzed showed a statistically significant association with reliability estimates, anyone of them achieved a nonsignificant result for the model misspecification test. Therefore, proposing a model containing the set of most relevant predictors of reliability was in order.

An explanatory model

With the purpose of finding a predictive model able to explain, at least, part of the variability in the reliability estimates, a weighted multiple regression analysis was applied assuming a mixed-effects model. Selection of predictors was based on both statistical and substantive criteria. Thus, we included in the model the following predictors: the standard deviation and the mean of test scores, the test version (0, original; 1, other), the mean age of the participants (in years), and the target population. Although from a conceptual point of view the predictive model was composed of five predictors, actually the model included seven independent variables, as the target population entered in the model as three dichotomous predictors enabling us to code, by applying a dummy coding system (0, absent; 1, present), the membership of the studies to nonclinical population, University students, and clinical population.

TABLE 4. Results of the weighted multiple regression analysis on the internal consistency reliability estimates (coefficients alpha + *KR*-21 estimates).

Moderator variable	b_j	Z_j	p
SD of test scores	-.395	-35.10	< .001
Mean of test scores	.045	12.91	< .001
Test version	-.036	-1.64	.101
Mean age (in years)	-.001	-.73	.464
Target population – Non clinical	-.022	-.31	.754
Target population – University	-.041	-.60	.551
Target population - Clinical	-.053	-.75	.453
Results of the multiple regression model:	$Q_R(7) = 1341.08$, $p < .001$; $R^2 = .979$		
	$Q_E(272) = 306.66$, $p = .073$		

Notes. b_j: partial unstandardized regression coefficients. Z_j: statistic test for testing the significance of each moderator variable once partialized out the influence of the other moderator variables in the model. p: probability level associated to Z_j. Q_R: statistic for testing the significance of the full regression model. Q_E: statistic for testing the model misspecification. R^2: proportion of variance accounted for by the full model.

Table 4 presents the results of the weighted multiple regression model. The full model reached a highly statistically significant result ($p < .001$) with a percentage of variance accounted for of 97.9%. Moreover, the model misspecification test was not statistically significant ($p = .073$). Although the model reached a large predictive power, the standard deviation and the mean of test scores were the only predictors that achieved a statistically significant contribution to the predictive model, the remaining predictors being unable to offer nothing relevant to the model. A result that deserves special attention is the absence of statistical significance for the target population. As Table 3 showed, this moderator variable presented statistically significant differences among the mean reliability estimates for the different target populations. However, the target population did not reach a statistically significant contribution to the predictive model. The reason of this result is the strong relationship between the target population and the mean of test scores. In particular, when the studies were classified as a function of the target population, the means obtained for the mean of test scores were 5.96 in nonclinical samples, 7.01 with University students, 11.40 for subclinical samples, and 13.11 for clinical samples. In fact, following the Botella *et al.*'s (2010) approach, a mixed-effects meta-analysis was carried out by taking the mean of test scores as the dependent variable and the target population as the independent variable. The results showed a statistically significant association between the mean of test scores and the target population [$Q_B(3) = 295.37$, $p < .001$; $R^2 = .29$]. Therefore, the reason for not finding a statistical contribution of the target population to the reliability estimates variability was the existence of a strong collinearity between the target population and the mean of test scores.

As only the standard deviation and the mean of test scores reached a statistically significant contribution to the predictive model, the definitive explanatory model was composed of these two predictors only. The mixed-effects multiple regression model for

these two predictors reached a statistically significant result [Q_R (2) = 1715.83, $p < .001$; $R^2 = .999$], with the two predictors exhibiting a significant contribution to the model ($p < .001$). Further, the misspecification test was non significant [Q_E (277) = 191.65, $p = .999$], meaning that the model was well-specified. As a consequence, the predictive equation $T' = .14 - 0.40*SD + .037*$Mean, can be used to obtain predictions of the reliability for a given standard deviation and mean of test scores. For example, taking the average of the *SD* (4.60) and of the mean (10.30) of test scores showed in Table 2, the equation gives a prediction for the reliability of test scores of .73.[3]

Discussion

As reliability is not a property of the test itself, but of the test scores obtained in each application, it is needed to develop RG studies in order to examine how reliability varies through different administrations of the same test. In this article we presented the results of an RG study about the MOCI, one of the most commonly applied tests in Psychology to assess obsessive and compulsive symptoms in psychiatric patients and as a screening tool for nonclinical population. As the MOCI is composed of dichotomous items, it was possible to take advantage of the *KR*-21 formula to increase the database of reliability estimates when the study did not report a coefficient alpha but reported the number of items of the test and the mean and standard deviation of test scores. Our comparison between coefficients alpha and *KR*-21 estimates showed a small underestimation of .02 with the *KR*-21 formula, very similar to that obtained by Lane *et al.* (2002). With this strategy it was possible to increase our database of internal consistency estimates from 44 to 308. Therefore, our recommendation is to use *KR*-21 formula in *RG* studies for tests composed of dichotomous items when the study does not report the coefficient alpha for the data at hand.

The mean reliability obtained for test scores of the total scale was .76, clearly over the cut off of .70 usually considered as the minimum recommendable reliability when applying tests for exploratory research purposes (Nunnally and Bernstein, 1994). The average test-retest reliability (.70) was slightly lower to that obtained for internal consistency, although still over .70. However, if we take into account the more strict cut off values of .80 for general research purposes and of .90 for important clinical decisions, then the reliability exhibited by the MOCI scores is clearly insufficient, as for clinical samples (see Table 3) the mean reliability coefficient was .75 only (Henson, 2001). Therefore, our results showed a satisfactory reliability for the MOCI only when it is used for exploratory research purposes.

Several characteristics of the studies presented a statistically significant relationship with the internal consistency estimates, such as the mean age, the target population, and the test version. However, when they were introduced in a multiple regression

[3] Note that the value obtained by applying the predictive equation was T' = -1.08. As this value is in the metric of Bonett's transformation, it is needed to back-transform this value in order to put the result in the metric of reliability coefficient. Thus, by applying the formula, we obtain .73.

model, the only variables that significantly contributed to explain the reliability estimates variability were the standard deviation and the mean of test scores. Therefore, our predictive model is based on only these two variables and it can be used in future research to predict the expected reliability in a given application of the MOCI.

Finally, only 15.6% of the studies that applied the MOCI reported a reliability coefficient with the data at hand. The majority of the studies that applied the MOCI induced reliability from previous applications of the test or simply omitted any comment about reliability. The erroneous practice of inducing reliability should be eradicated, as it does a disservice to psychological research (Vacha-Haase and Thompson, 2011).

References

Aguayo, R., Vargas, C., de la Fuente, E.I., and Lozano, L.M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology, 11*, 343-361.

Anastasi, A. and Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Ávila-Espada, A. (1986). Una contribución a la evaluación de las obsesiones y compulsiones con la revisión del Inventario de Obsesiones de Leyton (LOI) y del Cuestionario Obsesivo Compulsivo de Maudsley (MOCQ). *Psiquis*, *100*, 67-74.

Barnes, L.L.B., Harp, D., and Jung, W.S. (2002). Reliability generalization of scores on the Spielberger State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 62*, 603-618.

Bonett, D.G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*, 335-340.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Botella, J. and Gambara, H. (2006). Doing and reporting a meta-analysis. *International Journal of Clinical and Health Psychology*, *6*, 425-440.

Botella, J., Suero, M., and Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15*, 386-397.

Caruso, J.C., Witkiewitz, K., Belcourt-Dittloff, A., and Gottlieb, J.D. (2001). Reliability scores from the Eysenck Personality Questionnaire: A reliability generalization study. *Educational and Psychological Measurement, 61*, 675-689.

Chan, D.W. (1990). The Maudsley Obsessional-Compulsive Inventory: A psychometric investigation on Chinese normal subjects. *Behaviour Research and Therapy, 28*, 413-420.

Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.

Einstein, D.A. and Menzies, R.G. (2006). Magical thinking in obsessive-compulsive disorder, panic disorder and the general community. *Behavioural and Cognitive Psychotherapy*, *34*, 351-357.

Emmelkamp, P.M.G., Kraaijkamp, H.J.M., and Van den Hout, M.A. (1999). Assessment of obsessive-compulsive disorder. *Behavior Modification*, *23*, 269-279.

Erol, N. and Savasir, I. (1989, June). *The Turkish version of the Maudsley Obsessional Compulsive Questionnaire.* Paper presented at the 2nd Regional Conference of the International Association for Cross-Cultural Psychology. Amsterdam, The Netherlands.

Fonseca-Pedrero, E., Lemos-Giráldez, S., Paíno-Piñeiro, M., Villazón-García, U., and Muñiz, J. (2010). Schizotypal traits, obsessive-compulsive symptoms, and social functioning in adolescents. *Comprehensive Psychiatry, 51*, 71-77.

Fonseca-Pedrero, E., Paíno-Piñeiro, M., and Lemos-Giráldez, S. (2007). La diversidad psicopedagógica en el aula: Evaluación de problemas emocionales y comportamentales. *Aula Abierta, 36*, 39-48.

Guilera, G., Gómez-Benito, J., Tomás, J., and Carreras, V. (2005). Fiabilidad y validez de la versión española del Maudsley Obsessive-Compulsive Inventory (MOCI). *Interpsiquis*.

Hantouche, E.G. and Guelfi, J.D. (1993). Auto-évaluation du trouble obsessionnel-compulsif: Adaptation et validation de deux outils psychométriques en version française. *L'Encéphale*, *19*, 241-248.

Henson, R.K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177-189.

Henson, R.K., Kogan, L.R., and Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. *Educational and Psychological Measurement, 61*, 404-420.

Henson, R.K. and Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting «reliability generalization» studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.

Hodgson, R.J. and Rachman, S. (1977). Obsessional-compulsive complaints. *Behaviour Research and Therapy, 15*, 389-395.

Hopkins, K.D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston: Allyn and Bacon.

Kieffer, K.M. and Reese, R.J. (2002). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement, 62*, 969-994.

Lane, G.G., White, A.E., and Henson, R.K. (2002). Expanding reliability generalization methods with KR-21 estimates: An RG study on the Coopersmith Self-esteem Inventory. *Educational and Psychological Measurement, 62*, 685-711.

Li, C.-S.R. and Chen, S.-H. (2007). Obsessive-compulsiveness and impulsivity in a non-clinical population of adolescent males and females. *Psychiatry Research, 149*, 129-138.

López-Pina, J.A., Sánchez-Meca, J., and Rosa-Alcázar, A.I. (2009). The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology, 9*, 143-159.

Meesters, Y. (1997). Twee dwang-vragenlijsten bij OCD-patiënten: de MOCI en de IDB [Two compulsion inventories for OCD patients: MOCI and IDB]. *Gedragstherapie*, *30*, 103-112.

Min, B. and Won, H. (1999). Reliability and validity of the Korean translations of Maudsley Obsessional-Compulsive Inventory and Padua Inventory. *The Korean Journal of Clinical Psychology*, *18*, 163-182.

Montero I. and León, O. (2007). A guide for naming research studies in Psychology. *Internacional Journal of Clinical and Health Psychology*, *4*, 505-520.

Müller, N., Putz, A., Kathmann, N., Lehle, R., Günther, W., and Straube, A. (1997). Characteristics of obsessive-compulsive symptoms in Tourette syndrome, obsessive-compulsive disorder, and Parkinson disease. *Psychiatry Research*, *70*, 105-114.

Nunnally, J.C. and Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Pedhazur, E.J. and Schmelkin, L.P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.

Rachman, S. and Hodgson, R.J. (1980). *Obsessions and compulsions*. Englewood Cliffs, NJ: Prentice Hall.

Raich, R.M. (1996). Evaluación del trastorno obsesivo-compulsivo. In G. Buela-Casal, V. Caballo, and J.C. Sierra (Eds.), *Manual de evaluación en psicología clínica y de la salud* (pp. 161-178). Madrid: Siglo XXI.

Rodriguez, M.C. and Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322.

Rosa-Alcázar, A.I., Sánchez-Meca, J., Gómez-Conesa, A., and Marín-Martínez, F. (2008). The psychological treatment of obsessive-compulsive disorder: A meta-analysis. *Clinical Psychology Review*, *28*, 1310-1325.

Sanavio, E. and Vidotto, G. (1985). The components of the Maudsley Obsessional-Compulsive Questionnaire. *Behaviour Research and Therapy*, *23*, 659-662.

Sánchez-Meca, J. and López-Pina, J.A. (2008). El enfoque meta-analítico de generalización de la fiabilidad. *Acción Psicológica*, *5*, 37-64.

Sánchez-Meca, J. and Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*, 31-48.

Shields, A.L. and Caruso, J.C. (2004). A reliability induction and reliability generalization study of the Cage Questionnaire. *Educational and Psychological Measurement, 64*, 254-270.

Smári, J., Bjarnason, B.A., Porleifsson, K., Sturludóttir, S., and Hafsteinsdóttir, P. (1994). *Étude psychométrique de la Liste des Pensées Obsédantes en version islandaise*. Poster presented at the 22nd Conference of the French Association of Cognitive-Behaviour Therapy, Paris (France).

Sternberger, L.G., and Burns, G.L. (1990). Compulsive activity and the Maudsley Obsessional-Compulsive Inventory: Psychometric properties of two measures of obsessive-compulsive disorder. *Behavior Therapy, 21*, 117-127.

Støylen, I.J., Larsen, S., and Kvale, G. (2000). The Maudsley Obsessional-Compulsive Inventory and OCD in a norwegian nonclinical sample. *Scandinavian Journal of Psychology*, *41*, 283-286.

Tadai, T., Nakamura, M., Okazaki, S., and Nakajima, T. (1995). The prevalence of obsessive-compulsive disorder in Japan: A study of students using the Maudsley Obsessional-Compulsive Inventory and DSM-III-R. *Psychiatry and Clinical Neurosciences*, *49*, 39-41.

Thordarson, D.S., Radomsky, A.S., Rachman, S., Shafran, R., Sawchuk, C.N., and Hakstian, A.R. (2004). The Vancouver Obsessional Compulsive Inventory (VOCI). *Behaviour Research and Therapy*, *42*, 1289-1314.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20.

Vacha-Haase, T., Kogan, L.R., and Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals. *Educational and Psychological Measurement, 60*, 509-522.

Vacha-Haase, T. and Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44*, 159-168.

Vassar, M. and Hale, W. (2009). Reliability reporting across studies using the Buss Durkee Hostility Inventory. *Journal of Interpersonal Violence, 24*, 20-37.

Yin, P. and Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*, 201-223.

Zohar, A.H., LaBuda, M., and Moschel-Ravid, O. (1995). Obsessive-compulsive behaviors and cognitive functioning: A study of compulsivity, frame shifting and Type A activity patterns in a normal population. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, *8*, 163-167.

Received May 17, 2011
Accepted June 16, 2011