


The Yale–Brown Obsessive Compulsive Scale: A Reliability Generalization Meta-Analysis

Assessment
2015, Vol. 22(5) 619–628
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191114551954
asm.sagepub.com


José Antonio López-Pina¹, Julio Sánchez-Meca¹, José Antonio López-López¹, Fulgencio Marín-Martínez¹, Rosa María Núñez-Núñez², Ana I. Rosa-Alcázar¹, Antonia Gómez-Conesa¹, and Josefa Ferrer-Requena¹

Abstract

The Yale–Brown Obsessive Compulsive Scale (Y-BOCS) is the most frequently applied test to assess obsessive compulsive symptoms. We conducted a reliability generalization meta-analysis on the Y-BOCS to estimate the average reliability, examine the variability among the reliability estimates, search for moderators, and propose a predictive model that researchers and clinicians can use to estimate the expected reliability of the Y-BOCS. We included studies where the Y-BOCS was applied to a sample of adults and reliability estimate was reported. Out of the 11,490 references located, 144 studies met the selection criteria. For the total scale, the mean reliability was 0.866 for coefficients alpha, 0.848 for test–retest correlations, and 0.922 for intraclass correlations. The moderator analyses led to a predictive model where the standard deviation of the total test and the target population (clinical vs. nonclinical) explained 38.6% of the total variability among coefficients alpha. Finally, clinical implications of the results are discussed.

Keywords

reliability generalization, meta-analysis, Y-BOCS, internal consistency, test–retest reliability, intraclass correlation

According to the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM-IV-TR*; American Psychiatric Association, 2010) obsessive–compulsive disorder (OCD) is characterized by obsessions and compulsions that interfere with a person’s normal life. Obsessions are intrusive ideas, images, or impulses that are rejected by people who suffer from OCD. Obsessive issues can be related to the risk of damage or danger, contamination, germs or chemical products, doubts, concerns about symmetry, checking, and so on. Compulsions, on the other hand, are repetitive and stereotyped behaviors that require specific rules to be followed, their goal being to reduce the distress caused by obsessions. Subjects who suffer from this disorder can attract the attention of other people because of their irrational and unadapted behavior. In addition, the disorder causes a great functional impairment and disability in the subject (Abramowitz, Taylor, & McKay, 2009; Taylor, 2011). Epidemiological studies offer changing prevalence rates, but always larger for women than for men. Thus, in an Australian sample Crino, Slade, and Andrews (2005) found 12-month prevalence rates of 0.6% and 0.7% for men and women, respectively, whereas Kessler, Petukhova, Sampson, Zaslavsky, and Wittchen (2012) found a 12-month prevalence of 1.2% in a U.S. sample. In a systematic review of prevalence rates, Somers, Goldner, Waraich, and Hsu

(2006) found 12-month prevalence rates of 0.31% and 0.5% for men and women, respectively. Lifetime prevalence rates for OCD also vary, with figures of 1% and 1.6% for men and women in the Somers et al. (2006) study, and of 1.6% and 3%, respectively, in the Kessler et al. (2012) study.

Out of the different measurement instruments developed to assess obsessive–compulsive symptoms, the Yale–Brown Obsessive Compulsive Scale (Y-BOCS) is the most frequently applied test in clinical settings as well as in non-clinical population with screening purposes (Goodman, Price, Rasmussen, Mazure, Delgado, et al., 1989; Goodman, Price, Rasmussen, Mazure, Fleischmann, et al., 1989). The Y-BOCS is a clinician-rated scale that assesses the presence and severity of obsessions and compulsions indexed to the past week. The original version of the Y-BOCS was developed for adults and was composed of 10 items; 5 of them intended to assess the severity of obsessions and the other 5

¹Universidad de Murcia, Murcia, Spain

²Universidad Miguel Hernández de Elche, Elche, Spain

Corresponding Author:

Julio Sánchez-Meca, Department of Basic Psychology and Methodology, Faculty of Psychology, Espinardo Campus, University of Murcia, Murcia 30100, Spain.
Email: jsmecca@um.es

addressing compulsions. All items have a Likert-type scale scored from 0 to 4, so that the test offers a total score by summing the 10 items, as well as specific scores for the obsession and compulsion subscales. In addition, the Y-BOCS assesses the severity of OCD for a list of 54 symptoms dichotomously scored (present vs. absent) in terms of time spent, interference, distress, resistance, and control.

Although the Y-BOCS was initially developed to assess adults with OCD, it has later been adapted to children and adolescents (Scahill et al., 1997) as well as to other psychiatric disorders where obsessions and compulsions play a relevant role (Hollander et al., 1998; Mazure, Halmi, Sunday, Romano, & Einhorn, 1994; Modell, Glaser, Mountz, Schmaltz, & Cyr, 1992; Monahan, Black, & Gabel, 1996; Phillips et al., 1997). Several self-report versions of the Y-BOCS have also been developed (Summerfeldt, Richter, Antony, & Swinson, 1999). The Y-BOCS for adults has been adapted to different languages and cultures. Thus, different versions of the Y-BOCS have been published in at least 16 languages (Arrindell, De Vlaming, Eisenhardt, Van Berkum, & Kwee, 2002; Bejerot, Ekselius, & Von Knorring, 1998; Dome et al., 2006; Ghassemzadeh, Bolhari, Briaschk, & Salavati, 2005; Gross-Isserof et al., 1996; Hou, Yen, Huang, Wang, & Yeh, 2010; Jaisooriya, Reddy, & Srinath, 2003; Jónsson, Hougaard, & Beenedsen, 2011; Koponen et al., 1997; Lyoo, Lee, Kim, Kong, & Kwon, 2001; Mollard, Cottraux, & Bouvard, 1989; Moritz et al., 2002; Nakajima et al., 1995; Ólafsson, Snorrason, & Smári, 2010; Pertusa et al., 2010; Raszka et al., 2009; Rosario-Campos et al., 2006; Sica et al., 2004; Solem, Hjemdal, Vogel, & Stiles, 2010; Tek et al., 1995).

Reliability is one of the most critical properties of the test scores. An adequate reliability of the test scores is crucial for the clinician to reach an accurate diagnosis. Moreover, a low reliability can decrease the statistical power of the significance tests employed by applied researchers (Wilkinson & APA Task Force on Statistical Inference, 1999). Therefore, when using a psychometric instrument, reliability is a prerequisite to achieve valid conclusions at both the clinical and research contexts (Nunnally & Bernstein, 1994).

Goodman, Price, Rasmussen, Mazure, Fleischmann, et al. (1989) administered the Y-BOCS to 40 OCD patients, finding an interrater reliability of 0.98 or more, and a mean internal consistency among 4 raters of 0.89. Up to our knowledge, more than 50 psychometric studies of the Y-BOCS for adults have been published (Anholt et al., 2010; Boyette et al., 2011; Deacon & Abramowitz, 2005; De Haan et al., 2006; Storch et al., 2005). Their results offer good internal consistency (alpha coefficients between 0.58 and 0.98), test-retest reliability (Pearson correlations between 0.61 and 0.97), interrater agreement (intraclass correlations between 0.63 and 0.99, and kappa coefficients between 0.73 and 1.00). Nonetheless, these studies also

evidence a clear variability in the reliability estimates depending on the composition and variability of the samples. In addition, it is not clear whether the large number of different adaptations of the Y-BOCS to other languages and cultures exhibit similar reliability estimates from the test scores.

When a test is applied to a sample of participants, researchers should report a reliability estimate with the data at hand. However, it is very common to find that researchers have induced score reliability from previous administrations of the test to other samples (Green, Chen, Helms, & Henze, 2011). Reliability induction is an erroneous practice because, as psychometric theory states, reliability is not a property of the test itself, but of the test scores (Crocker & Algina, 1986; Lord & Novick, 1968; McDonald, 1999; Streiner & Norman, 2008). Therefore, reliability should be determined using the sample in which the Y-BOCS is administered.

As score reliability changes from a test administration to the next, the best way to guide expectations about the reliability of the test scores is to quantitatively integrate several reliability estimates obtained from different administrations of the instrument. To this respect, meta-analysis constitutes an optimal method to examine how score reliability varies along different test applications. In this vein, Vacha-Haase (1998) coined the term *reliability generalization* (RG) to refer to this kind of meta-analysis. In an RG meta-analysis, an exhaustive search of the studies that have applied the test is carried out, and those that report any reliability estimate with the own sample data are included in the meta-analysis (Henson & Thompson, 2002; Rodriguez & Maeda, 2006; Sánchez-Meca, López-López, & López-Pina, 2013; Vacha-Haase & Thompson, 2011).

Objectives

We conducted an RG meta-analysis on the Y-BOCS for adults (a) to estimate the average reliability, in terms of internal consistency, test-retest reliability, and interrater agreement, found in the empirical studies that applied the Y-BOCS; (b) to examine the variability among the reliability estimates; (c) if there is more variability than sampling error can explain, to search for substantive and methodological characteristics of the studies that can be statistically associated to the reliability coefficients; and (d) to propose a predictive model that researchers and clinicians can use in the future to estimate the expected reliability of the Y-BOCS as a function of the most relevant study characteristics (Henson & Thompson, 2002; Rodriguez & Maeda, 2006). In particular, it was expected that characteristics such as the mean and the standard deviation of the test scores, the mean age, the target population of the participants (clinical vs. nonclinical), and the test version (original vs. adapted), would affect the score reliability.

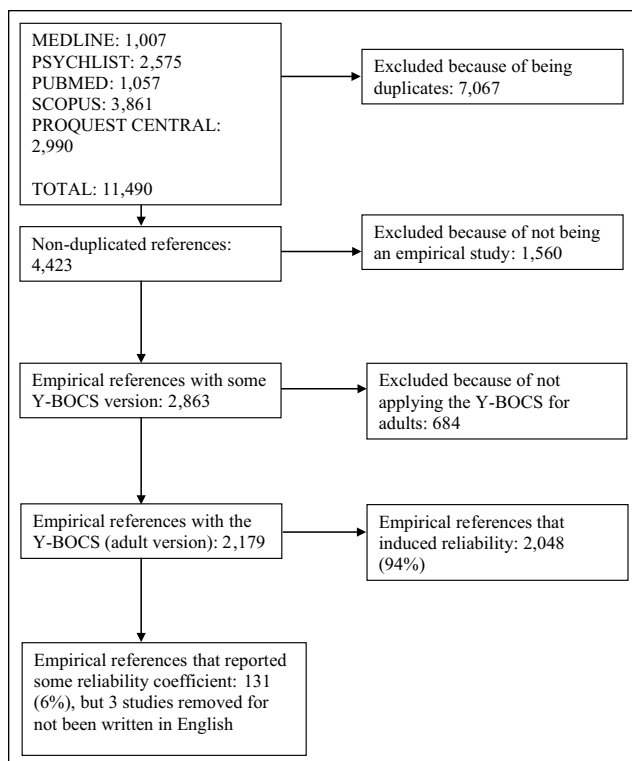


Figure 1. Flowchart describing the search strategy to select the studies for the meta-analysis.

Method

Data Sources

As the Y-BOCS was developed in 1989, the search period of the relevant studies covered from 1989 to 2011 inclusive. The following databases were consulted: MedLine, SCOPUS, PsycInfo, PUBMED, and PROQUEST CENTRAL. In the search, the following keywords were combined to be found along the documents: “Yale–Brown obsessive compulsive scale” or “Y-BOCS” or “YBOCS” or “YBOC.” A complementary electronic search was launched combining the keywords: “Factor analysis” and “Y-BOCS” or “Reliability” and “Y-BOCS” or “Validity” and “Y-BOCS.”

Study Selection

To be included in the meta-analysis, the study had to comply with three criteria: (a) to be an empirical study where the Y-BOCS was applied to an adult sample (≥ 18 years old), (b) to report any reliability estimate with data from the indexed (study) sample, and (c) it had to be written in English.

Figure 1 presents a flowchart describing the selection process of the studies. The search yielded a total of 11,490 references, out of which 9,311 were removed for different

reasons. The remaining 2,179 references were empirical studies that had applied the Y-BOCS for adults. Out of these, 131 (6%) studies reported any estimate of the test scores reliability (although 3 studies were removed from our analyses because they were not written in English), whereas the remaining 2,048 (94%) induced reliability from other studies. Two kinds of reliability induction can be distinguished (Shields & Caruso, 2004): Reliability induction by omission consists of omitting any reference to the test score reliability, whereas reliability induction by report occurs when the study reports some reliability estimate from previous studies. Out of the 2,048 studies that induced reliability, 1,734 (84.7%) omitted any reference to the Y-BOCS reliability, whereas the remaining 314 studies (15.3%) induced reliability by reporting a previous reliability estimate. Therefore, the database of our RG meta-analysis was based on the 128 studies that reported any reliability estimate with the data at hand.

Data Extraction

To explore how study characteristics can affect score reliability when the Y-BOCS is applied, the following moderator variables were coded in the studies: (a) standard deviation of the total test scores; (b) mean of the total test scores; (c) test version (original vs. other); (d) administration format (clinical interview vs. self-administered); (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) target population of the sample (nonclinical vs. clinical); (i) disorder of the participants (OCD vs. other); (j) mean of disorder history (in years); (k) standard deviation of disorder history (in years); (l) study focus (psychometric vs. substantive); (m) focus of the psychometric studies (Y-BOCS vs. other tests); (n) publication year; (o) continent (Europe, North America, South America, Asia, or Oceania); (p) main researcher affiliation (psychology vs. psychiatry), and (q) sample size. Together with these moderator variables, coefficients alpha, test–retest, and intraclass correlations, as well as other types of reliability estimate, 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 Y-BOCS. This sample of studies was doubly coded by two independent coding teams, whose members were psychologists with a PhD in psychology and specialized in meta-analysis. In general, the inter-coder agreement was satisfactory, with kappa coefficients ranging between 0.65 and 0.99 for the qualitative characteristics, and intraclass correlations ranging between 0.69 and 1.00 for the continuous variables. The inconsistencies between the coders were solved by consensus.

Data Synthesis

Separate meta-analyses were conducted for coefficients alpha, test–retest, and intraclass correlations, as they estimate three different types of reliability, following the recommendations of several authors from the RG meta-analytic arena (Dimitrov, 2002; Sawilowski, 2000). Several transformations were applied on the coefficients in order to normalize their sampling distributions and to stabilize their variances (Botella, Suero, & Gambará, 2010; Rodríguez & Maeda, 2006; Sánchez-Meca et al., 2013; Shields & Caruso, 2004; Vacha-Haase & Thompson, 2011). Coefficients alpha were transformed 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. On the other hand, test–retest and intraclass correlations were transformed using Fisher's Z (Sánchez-Meca et al., 2013).

To obtain summary statistics of reliability coefficients, random-effects models were assumed and, consequently, the reliability coefficients were weighted by the inverse variance defined as the sum of the within-study and the between-studies variances. The latter was estimated using the empirical Bayes method (López-López, Marín-Martínez, Sánchez-Meca, Van den Noortgate, & Viechtbauer, 2014; Morris, 1983). The confidence limits around the overall reliability estimates were computed using the method proposed by Hartung (1999, Sánchez-Meca & Marín-Martínez, 2008). The heterogeneity exhibited by the reliability estimates was assessed with the Q test and the I^2 index.

Finally, the moderator analyses were conducted through regression analyses for the continuous variables and analyses of variance for the qualitative ones. The analyses assumed mixed-effects models and used the adjustment proposed by Knapp and Hartung (2003; López-López, Botella, Sánchez-Meca, & Marín-Martínez, 2013).

To facilitate interpretation of results, the average reliability estimates, their confidence limits, and the slope estimates were back-transformed to the original metric of reliability coefficients. The different formulas employed for such back-transformations can be found elsewhere (López-López et al., 2013; Sánchez-Meca et al., 2013).

Last, some sensitivity analyses were conducted. The analyses detailed above using the transformed reliability coefficients were also carried out using the untransformed coefficients for comparison purposes. Moreover, the risk of publication bias was assessed constructing funnel plots and applying the trim-and-fill method (Duval & Tweedie, 2000). All statistical analyses were carried out with the metafor package in R (Viechtbauer, 2010).

Results

Descriptive Characteristics of the Studies

The present RG meta-analysis was focused on the 128 studies written in English that reported any reliability estimate.

Because of space limitations, the list of study references is not reported in the article, but it can be obtained from the corresponding author on request. Regarding the location of the studies, 27.3% were conducted in Europe, 59.4% in North America, 10.2% in Asia, 1.6% in Oceania, and the remaining 1.5% corresponded to a Turkish and a multi-center study.

Despite 128 articles reported at least one reliability estimate, our unit of analysis was the sample. Therefore, given that several studies reported more than one reliability coefficient for different samples, we collected a total of 235 reliability estimates from 144 independent samples. The most frequently reported reliability estimate was coefficient alpha, computed from 79 (54.9%) different samples, leading to a pooled sample of $N = 11,512$ participants. Other types of reliability found were inter-rater agreement coefficients, with the intraclass correlation reported in 41 (28.5%) samples ($N = 2,650$) and the kappa coefficient computed for 12 (8.3%) samples ($N = 414$). Also, 13 test–retest correlations were retrieved (9% samples, $N = 741$). Regarding the obsessions and compulsions subscales, coefficient alpha was reported in 31 (21.5%) samples ($N = 3,848$), the intraclass correlation was computed for 8 (5.5%) samples ($N = 259$), and 5 test–retest correlations were reported (3.5% samples, $N = 124$).

Mean Reliability and Heterogeneity

Table 1 shows the main summary statistics for coefficients alpha. The 79 estimates reported for the total scale yielded a (weighted) mean coefficient alpha of 0.866 (95% confidence limits: 0.849 and 0.882). For the subscales, coefficients alpha were computed for 31 different samples, leading to an overall estimate of 0.824 (confidence limits: 0.789 and 0.854) for the obsessions subscale, and an average coefficient of 0.837 (limits: 0.806 and 0.862) for the compulsions subscale. Table 1 also presents the results of the Q statistics and the I^2 indices for the assessment of the variability exhibited by the reliability estimates. Coefficients alpha for the total scale and subscales showed a statistically significant heterogeneity, with I^2 values around 90%. Note that I^2 values of 25%, 50%, and 75% can be interpreted as reflecting small, medium, and large heterogeneity among the reliability coefficients, respectively (Higgins & Thompson, 2002). Consequently, analyses to explain part of that heterogeneity were in order.

Regarding test–retest reliability, the 13 correlations computed for the total scale led to an average estimate of 0.848 (confidence limits: 0.772 and 0.900) for the total scale. Also, 5 samples reported test–retest correlations for the subscales, and the overall reliability estimate was 0.725 (confidence limits: 0.437 and 0.879) for the obsessions subscale and 0.673 (confidence limits: 0.472 and 0.807) for the subscale of compulsions. Significant heterogeneity was found only for the total scale, with an I^2 value of 56.86%.

Table 1. Overall Reliability and 95% Confidence Intervals for the Alpha Coefficients, Test–Retest, and Intraclass Correlations of the Total Scale and the Obsessions and Compulsions Subscales.

Scale/Subscale	<i>k</i>	Minimum	Maximum	Mean	95% CI [Lb, Ub]	<i>Q</i>	<i>I</i> ²
Coefficients α							
Total scale	79	0.58	0.98	0.866	[0.849, 0.882]	1985.548***	94.55
Obsessions	31	0.55	0.97	0.824	[0.789, 0.854]	222.299***	91.29
Compulsions	31	-0.34	0.95	0.837	[0.806, 0.862]	182.148***	89.91
Test–retest							
Total scale	13	0.61	0.97	0.848	[0.772, 0.900]	27.839**	56.86
Obsessions	5	0.55	0.92	0.725	[0.437, 0.879]	4.653	15.40
Compulsions	5	0.52	0.82	0.673	[0.472, 0.807]	2.319	0
Intraclass correlation							
Total scale	41	0.63	0.99	0.922	[0.896, 0.941]	286.285***	78.35
Obsessions	8	0.72	0.97	0.936	[0.880, 0.967]	11.941	42.81
Compulsions	8	0.64	0.96	0.927	[0.866, 0.961]	10.618	38.02

Note. *k* = number of studies (or reliability coefficients); Lb and Ub = lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; *Q* = heterogeneity statistic; *I*² = heterogeneity index. In order to facilitate the interpretation, all means and their respective confidential limits were back-transformed to the metric of the original coefficients.

p* < .05. *p* < .01. ****p* < .001.

With respect to interrater agreement, intraclass correlations for the total scale were computed from 41 samples, yielding an average reliability of 0.922 (confidence limits: 0.896 and 0.941). Moreover, 8 correlations were collected for the subscales, leading to an overall reliability of 0.936 (confidence limits: 0.880 and 0.967) for the obsessions subscale and 0.927 (confidence limits: 0.866 and 0.961) for the compulsions subscale. Significant heterogeneity was found only among the intraclass correlations from the total scale, with an *I*² value of 78.35%. Nonetheless, the *I*² values obtained for the obsessions and compulsions subscales (42.81% and 38.02%) suggested some heterogeneity as well.

Moderator Analyses

As coefficient alpha was the most frequently reported reliability estimate, moderator analyses were conducted only for these coefficients. Table 2 shows the results of the weighted simple regression analyses conducted for the continuous moderators, with the transformed coefficients alpha of the total scale as the dependent variable. As the psychometric theory predicts, there was a positive, statistically significant relationship between the standard deviation of the total test scores and the reliability estimates (*p* < .001) with a 37.9% of variance accounted for. In addition, a negative, statistically significant relationship was found between the mean of the test scores and the reliability estimates (*p* = .001) with a 17.1% of variance accounted for. The sample size showed a significant relationship with the coefficients as well (*p* = .020), although the *R*_{adj}² index yielded only a 6.3% of variance explained. The standard deviation of the age of the participants also showed a negative, statistically

significant relationship with the reliability coefficients (*p* = .005), with a 13.3% of variance accounted for. The gender distribution in the samples also showed a statistically significant relationship with coefficients alpha (*p* = .037), with reliability estimates increasing as the percentage of males in the samples decreased. However, gender distribution only explained a 6% of the coefficients alpha's variability. Last, the mean of the disorder history also achieved a negative, significant result (*p* = .023), with a 42.2% of variance accounted for. The remaining continuous moderator variables here tested did not reach the statistical significance: mean of the participants age, publication year, and the standard deviation of the disorder history.

With regard to the qualitative moderators, Table 3 shows the results of the weighted analyses of variance applied on the coefficients alpha of the total scale. The target population showed a statistically significant influence on the reliability estimates (*p* = .001) and a 14.1% of variance explained, with a higher overall reliability for the nonclinical samples ($\hat{\alpha}_+ = 0.904$) than for the clinical samples ($\hat{\alpha}_+ = 0.848$). For the clinical samples, the type of disorder was associated with the heterogeneity among the coefficients as well (*p* < .001), accounting for 22.7% of that variability and showing a higher average reliability for non-OCD samples. Out of the 82 studies that reported a coefficient alpha, 35 of them were psychometric studies. When those 35 studies were classified as a function of whether the target test had been the Y-BOCS or another one, statistically significant differences were found between their mean coefficients alpha (*p* = .036), with a 10.8% of variance explained, the mean reliability being lower for psychometric studies focused on the Y-BOCS. Last, as Table 3 shows, there were no statistically significant

Table 2. Results of the Simple Meta-Regression Analyses Assuming a Mixed-Effects Model on the Transformed Alpha Coefficients for the Continuous Moderator Variables.

Moderator variable	k	b_j	t	p	R_{adj}^2	Q_E
SD of the total scores	63	0.081	-5.676	<.001	.379	503.69***
Mean of the total scores	64	-0.002	3.492	.001	.171	619.79***
Sample size	79	0.000	-2.376	.020	.063	1263.25***
Mean age (in years)	59	-0.001	1.514	.136	.024	1518.87***
SD of the age (in years)	56	-0.005	2.931	.005	.133	1298.96***
Percentage of males in the sample	63	-0.001	2.130	.037	.060	1851.50***
Year of publication	79	251.23	-0.554	.581	0	1976.61***
Mean of disorder history (in years)	13	-0.005	2.651	.023	.422	35.39***
SD of disorder history (in years)	12	-0.008	2.133	.059	.304	43.80***

Note. k = number of studies; b_j = unstandardized regression coefficient; t = significance test of the regression coefficient; p = p value of the significance test. R_{adj}^2 = proportion of variance explained; Q_E = statistic to test the model misspecification. In order to facilitate the interpretation, the regression coefficients were back-transformed to the metric of the original coefficients.

* p < .05. ** p < .01. *** p < .001.

Table 3. Results of the Weighted ANOVAs Assuming a Mixed-Effects Model on the Transformed Alpha Coefficients for the Categorical Moderator Variables.

Moderator variable	k_j	$\bar{\alpha}_j$	95% CI [α_l, α_u]	ANOVA results
Test version				$Q_B = 0.608, p = .438$
Original	54	0.871	[0.850, 0.889]	$R_{adj}^2 = 0$
Adapted	25	0.856	[0.820, 0.885]	$Q_W = 1855.94, p < .001$
Administration format				$Q_B = 3.093, p = .083$
Clinical interview	48	0.853	[0.828, 0.875]	$R_{adj}^2 = .029$
Self-administered	31	0.883	[0.858, 0.903]	$Q_W = 1891.53, p < .001$
Study focus				$Q_B = 0.410, p = .524$
Psychometric	35	0.872	[0.846, 0.894]	$R_{adj}^2 = 0$
Substantive	44	0.861	[0.836, 0.883]	$Q_W = 1814.16, p < .001$
Psychometric focus				$Q_B = 4.757, p = .036$
Y-BOCS	26	0.854	[0.816, 0.885]	$R_{adj}^2 = .108$
Other	9	0.910	[0.868, 0.939]	$Q_W = 805.20, p < .001$
Continent				$Q_B = 1.019, p = .366$
Europe	16	0.846	[0.797, 0.883]	$R_{adj}^2 = .001$
North America	52	0.872	[0.851, 0.890]	$Q_W = 1668.69, p < .001$
Asia	9	0.886	[0.833, 0.923]	
Target population				$Q_B = 12.30, p = .001$
Nonclinical	21	0.904	[0.880, 0.922]	$R_{adj}^2 = .141$
Clinical	58	0.848	[0.825, 0.868]	$Q_W = 1476.15, p < .001$
Researcher affiliation				$Q_B = 0.144, p = .706$
Psychologist	34	0.862	[0.835, 0.884]	$R_{adj}^2 = 0$
Psychiatrist	31	0.855	[0.825, 0.880]	$Q_W = 1502.97, p < .001$
Disorder				$Q_B = 16.61, p < .001$
OCD	55	0.847	[0.822, 0.868]	$R_{adj}^2 = .227$
Other	6	0.940	[0.907, 0.961]	$Q_W = 735.23, p < .001$

Note. ANOVA = analysis of variance; Y-BOCS = Yale-Brown Obsessive Compulsive Scale; k_j = number of studies (or coefficients) for each category of the moderator variable; $\bar{\alpha}_j$ = Average reliability coefficient for each category of the moderator variable. α_l and α_u = lower and upper confidence limits, respectively, for each average reliability coefficient; Q_B = between-categories homogeneity test; p = p value for the statistical tests; R_{adj}^2 = proportion of variance explained; Q_W = Within-category statistic for testing the model misspecification; OCD = obsessive-compulsive disorder. In order to facilitate the interpretation, the average reliability coefficients and their respective confidence limits were back-transformed to the metric of the original coefficients.

differences when the samples were classified as a function of the test version administered (original vs. adapted), the administration format (interview vs. self-report), study focus (psychometric vs. substantive), continent (Europe, North America, or Asia, after discarding other areas from which less than 5 studies were retrieved), or by the main researcher affiliation (psychology vs. psychiatry).

Although some of the moderators showed a statistically significant association with the reliability coefficients, none of them achieved a nonsignificant result for the model misspecification test (Q_E or Q_W for continuous and qualitative moderators, respectively), which suggested the presence of residual heterogeneity among the reliability coefficients after including the moderator (Hedges & Olkin, 1985). Therefore, a last objective of this meta-analysis was to propose an explanatory model containing the set of most relevant predictors of the coefficients alpha.

An Explanatory Model

With the aim to find a predictive model able to explain, at least, part of the variability among the reliability estimates, a weighted multiple meta-regression analysis was applied assuming a mixed-effects model. From both a statistical and substantive basis, two moderator variables were included in the model: the standard deviation of total test scores and the target population (0 = *nonclinical*; 1 = *clinical*). Although other moderator variables exhibited a statistical relationship with coefficient alpha, they were not included in the predictive model due to the presence of missing data. Once applied the multiple meta-regression analysis, the model coefficients were back-transformed to the metric of the coefficients, leading to the following equation:

$$\hat{\alpha}_i = 0.774 + 0.018SD_i - 0.038POPULATION_i$$

The full model reached a statistically significant result ($p < .001$) with a 38.6% of variance accounted for. When testing individually the predictors, both the standard deviation and the target population (nonclinical vs. clinical) showed a statistically significant relationship with the coefficients alpha ($p < .001$ and $p = .016$, respectively), with higher reliability predictions for a higher standard deviation ($b_1 = 0.018$) and a nonclinical sample ($b_2 = -0.038$). As a counterpart, the model misspecification test was also statistically significant ($p < .001$), therefore suggesting that other study characteristics were affecting the coefficients alpha variability as well.

Sensitivity Analyses

To check the robustness of our results, the analyses were repeated using the untransformed coefficients alpha, test-retest, and intraclass correlations. The analyses conducted

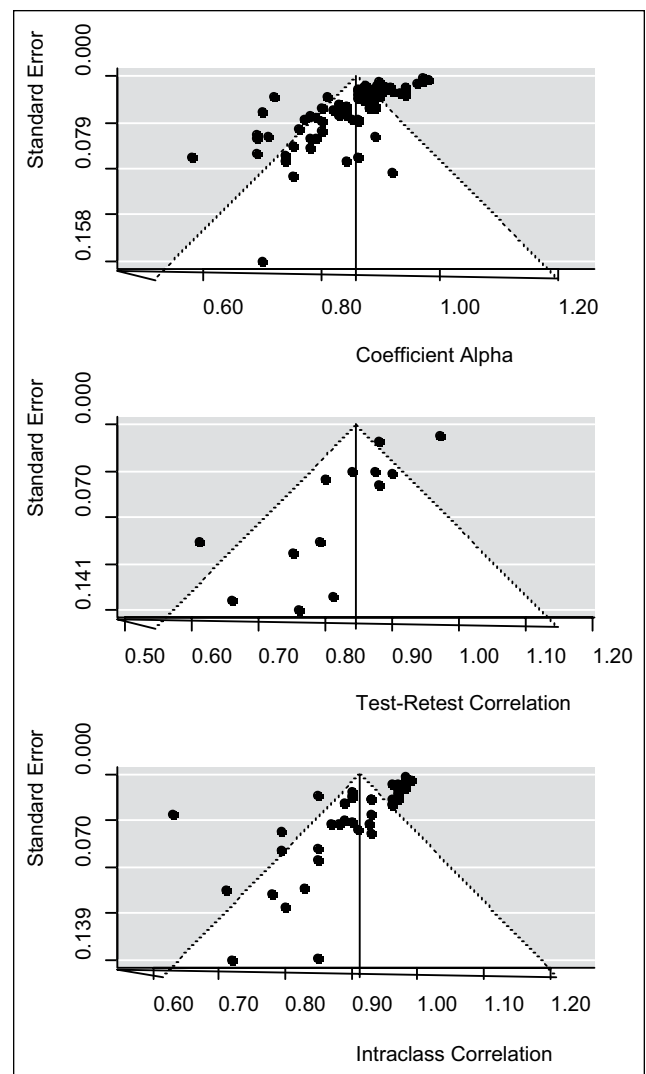


Figure 2. Funnel plots of the reliability estimates for the total Yale-Brown Obsessive Compulsive Scale (Y-BOCS).

with the untransformed coefficients did not show important discrepancies compared with the results presented along this section. On the other hand, the presence of publication bias in our results was checked by constructing funnel plots and applying the trim-and-fill method (Duval & Tweedie, 2000). Figure 2 presents the funnel plots obtained for coefficients alpha, test-retest, and intraclass correlations. When the trim-and-fill method was applied on each funnel plot, no coefficients were imputed in the left hand of the graph. Therefore, publication bias can be discarded as a threat to the meta-analytic results.

Discussion

As reliability is not a property of the test itself, but of the test scores obtained in each application, RG studies allow

one to examine how reliability varies through different test administrations, and to guide reliability expectations in future test applications. In this article, we presented the results of an RG study about the Y-BOCS, the most commonly applied test for the assessment of obsessive and compulsive symptoms in psychiatric patients and for the screening of nonclinical population.

The most commonly reported reliability estimate was coefficient alpha, with a weighted average for the total scale of 0.866. Test-retest and intraclass reliability exhibited means of 0.848 and 0.922, respectively. Thus, on average, the three types of reliability were clearly over the cutoff of 0.70, usually considered as the minimum recommended reliability when a test is administered with exploratory research purposes. The results are also satisfactory when taking the limit of 0.80 recommended for general research purposes (Nunnally & Bernstein, 1994). However, considering the stricter criterion of 0.90 determined when important clinical decisions are derived from the test scores, only intraclass correlation provided appropriate reliability estimates on average. These results pose into question the general adequacy of the Y-BOCS in terms of its reliability when this instrument is administered with clinical purposes, especially because the mean coefficient alpha was under 0.90 for clinical samples (0.848, see Table 3). The results of this RG meta-analysis suggest, therefore, that the Y-BOCS not only provides consistent information for its use with research purposes but also that the scores should be interpreted cautiously when this instrument is applied at a clinical context.

Our results also showed a large variability among the reliability estimates. Several characteristics of the studies presented a statistically significant relationship with the coefficients alpha. The multiple regression model including the standard deviation of the total test scores and the target population accounted for 38.6% of the heterogeneity among the coefficients, revealing that the highest reliability estimates can be expected when the variability among the test scores is large and with nonclinical samples.

Finally, it is worth noting that only 6% of the studies that applied the Y-BOCS computed a reliability coefficient with the data at hand. The remaining studies either induced reliability from previous applications or did not even mention reliability along the text (see Figure 1). This incorrect practice of inducing reliability not only affects the Y-BOCS but also any measurement instrument used in the psychiatric and psychological research (Green et al., 2011; Thompson, 2003). Researchers and practitioners must be aware that reliability is not a stable property of the test, so that it should always be estimated when a psychometric instrument is administered (Nunnally & Bernstein, 1994; Vacha-Haase, 1998; Wilkinson & Task Force on Statistical Inference, 1999). Meta-analytic RG studies are needed to detect the problem of reliability induction and to raise awareness

among clinicians and researchers about the importance of reporting reliability estimates with the own sample data.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by a grant from the Fundacion Seneca, Region of Murcia (Spain; Project No 08650/PHCS/08).

References

- Abramowitz, J. S., Taylor, S., & McKay, D. (2009). Obsessive compulsive disorder. *Lancet*, *374*, 491-499.
- American Psychiatric Association. (2010). *DSM-5 development*. Washington, DC: Author. Retrieved from <http://www.dsm5.org/Pages/Default.aspx>
- Anholt, G. E., Cath, D. C., Van Oppen, P., Eikelenboom, M., Smith, J. H., Van Meggen, H., . . . Van Balkon, A. J. L. M. (2010). Autism and ADHD symptoms in patients with OCD: Are they associated with specific OC symptom dimensions or OC symptom severity? *Journal of Autism and Developmental Disorders*, *40*, 580-589.
- Arrindell, W. A., De Vlaming, I. H., Eisenhardt, M. B., Van Berkum, D. E., & Kwee, M. G. T. (2002). Cross-cultural validity of the Yale-Brown Obsessive-Compulsive Scale. *Journal of Behavior Therapy and Experimental Psychiatry*, *33*, 159-176.
- Bejerot, S., Ekselius, L., & Von Knorring, L. (1998). Comorbidity between obsessive-compulsive disorder (OCD) and personality disorders. *Acta Psychiatrica Scandinavica*, *97*, 398-402.
- Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*, 335-340.
- Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods*, *15*, 386-397.
- Boyette, L., Swets, M., Meijer, C., Wouters, L., & G.R.O.U.P. authors. (2011). Factor structure of the Yale-Brown Obsessive Compulsive Scale (Y-BOCS) in a sample of patients with schizophrenia or related disorders and comorbid obsessive-compulsive symptoms. *Psychiatry Research*, *186*, 409-413.
- Crino, R., Slade, T., & Andrews, G. (2005). The changing prevalence and severity of obsessive-compulsive disorder criteria from *DSM-III* to *DSM-IV*. *American Journal of Psychiatry*, *162*, 876-882.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart and Winston.
- De Haan, L., Hooegeboom, B., Beuk, N., Wouters, L., Dingemans, P. M., & Linszen, D. H. (2006). Reliability and validity of the Yale-Brown Obsessive Compulsive Scale in Schizophrenia patients. *Psychopharmacology Bulletin*, *39*, 25-40.
- Deacon, B. J., & Abramowitz, J. S. (2005). The Yale-Brown Obsessive Compulsive Scale: Factor analysis, construct

- validity, and suggestions for refinement. *Journal of Anxiety Disorders*, 19, 573-585.
- Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*, 62, 783-801.
- Dome, P., Teleki, Z., Gonda, X., Gaszner, G., Mandl, P., & Rihmer, Z. (2006). Relationship between obsessive-compulsive symptoms and smoking habits amongst schizophrenic patients. *Psychiatry Research*, 14, 227-231.
- Duval, S. J., & Tweedie, R. L. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455-463.
- Ghassemzadeh, H., Bolhari, J., Briaschk, B., & Salavati, M. (2005). Responsibility attitude in a sample of Iranian obsessive-compulsive patients. *International Journal of Social Psychiatry*, 51, 13-22.
- Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Delgado, P., Heninger, G. R., & Charney, D. S. (1989). The Yale-Brown Obsessive Compulsive Scale II: Validity. *Archives of General Psychiatry*, 46, 1012-1016.
- Goodman, W. K., Price, L. H., Rasmussen, S. A., Mazure, C., Fleischmann, R. L., Hill, C. L., . . . Charney, S. (1989). The Yale-Brown Obsessive Compulsive Scale I. Development, use, and reliability. *Archives of General Psychiatry*, 46, 1006-1011.
- Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in *Psychological Assessment*: Recognizing the people behind the data. *Psychological Assessment*, 23, 656-669.
- Gross-Isserof, R., Sasson, Y., Voet, H., Hendler, T., Luca-Haimovici, K., Kandel-Sussman, H., & Zohar, J. (1996). Alternation learning in obsessive-compulsive disorder. *Biological Psychiatry*, 39, 733-738.
- Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, 41, 901-916.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting «reliability generalization» studies. *Measuring and Evaluation in Counseling and Development*, 35, 113-127.
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539-1558.
- Hollander, E., DeCaria, C. M., Mari, E., Wong, C. M., Mosovich, S., Grossman, R., & Begaz, T. (1998). Short-term single-blind fluvoxamine treatment of pathological gambling. *American Journal of Psychiatry*, 155, 1781-1783.
- Hou, S-Y., Yen, C-F., Huang, M-F., Wang, P-W., & Yeh, Y-C. (2010). Quality of Life and its correlates in patients with obsessive-compulsive disorder. *Kaohsiung Journal of Medical Sciences*, 26, 397-407.
- Jaisooriya, T. S., Reddy, J., & Srinath, S. (2003). The relationship of obsessive-compulsive disorder to putative spectrum disorders: Results from an Indian study. *Comprehensive Psychiatry*, 44, 317-323.
- Jónsson, H., Hougaard, E., & Beenedsen, B. E. (2011). Randomized comparative study of group versus individual cognitive behavioural therapy for obsessive compulsive disorder. *Acta Psychiatrica Scandinavica*, 123, 387-397.
- Kessler, R. C., Petukhova, M., Sampson, N. A., Zaslavsky, A. M., & Wittchen, H-U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International Journal of Methods in Psychiatric Research*, 21, 169-184.
- Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22, 2693-2710.
- Koponen, H., Lepola, U., Leinonen, E., Jokinen, R., Penttinen, J., & Turtonen, J. (1997). Citalopram in the treatment of obsessive-compulsive disorder: An open pilot study. *Acta Psychiatrica Scandinavica*, 96, 343-346.
- López-López, J. A., Botella, J., Sánchez-Meca, J., & Marín-Martínez, F. (2013). Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *Journal of Educational and Behavioral Statistics*, 38, 443-469.
- López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology*, 67, 30-48.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.
- Lyoo, I. K., Lee, D. W., Kim, Y. S., Kong, S. W., & Kwon, J. S. (2001). Patterns of temperament and character in subjects with obsessive-compulsive disorder. *Journal of Clinical Psychiatry*, 62, 637-641.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.
- Mazure, C. M., Halmi, K. A., Sunday, S. R., Romano, S. J., & Einhorn, A. M. (1994). The Yale-Brown-Cornell eating disorder scale: Development, use, reliability and validity. *Journal of Psychiatry Research*, 28, 425-445.
- Modell, J. G., Glaser, F. B., Mountz, J. M., Schmaltz, S., & Cyr, L. (1992). Obsessive and compulsive characteristics of alcohol abuse and dependence: Quantification by a newly developed questionnaire. *Alcoholism, Clinical and Experimental Research*, 16, 266-271.
- Mollard, E., Cottraux, J., & Bouvard, M. (1989). Version française de l'Échelle d'obsession-compulsion de Yale-Brown [French version of the Yale-Brown Obsessive Compulsive Scale]. *L'Encéphale*, XV, 335-341.
- Monahan, P., Black, D. W., & Gabel, J. (1996). Reliability and validity of a scale to measure change in persons with compulsive buying. *Psychiatry Research*, 64, 59-67.
- Moritz, S., Meier, B., Kloss, M., Jacobsen, D., Wein, C., Fricke, S., & Hand, I. (2002). Dimensional structure of the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS). *Psychiatry Research*, 109, 193-199.
- Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of American Statistical Association*, 78, 47-55.
- Nakajima, T., Nakamura, M., Taga, C., Yamagami, S., Kiririke, N., Nagata, T., . . . Yamaguchi, K. (1995). Reliability and validity of the Japanese version of the Yale-Brown Obsessive-Compulsive Scale. *Psychiatry and Clinical Neurosciences*, 49, 121-126.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

- Ólafsson, R. P., Snorrason, Í., & Smári, J. (2010). Yale-Brown Obsessive Compulsive Scale: Psychometric properties of the self-report version in a student sample. *Journal of Psychopathology and Behavioral Assessment*, 32, 226-245.
- Pertusa, A., Jaurrieta, N., Real, E., Alonso, P., Bueno, B., Segalás, C., . . . Menchón, J. M. (2010). Spanish adaptation of the Dimensional Yale-Brown Obsessive-Compulsive Scale. *Comprehensive Psychiatry*, 51, 641-648.
- Phillips, K. A., Hollander, E., Rasmussen, S. A., Aronowitz, B. R., Decaria, C., & Goodman, W. K. (1997). A severity rating scale for body dysmorphic disorder: Development, reliability, and validity of a modified version of the Yale-Brown Obsessive-Compulsive Scale. *Psychopharmacology Bulletin*, 33, 17-22.
- Raszka, M., Prasko, J., Koprivova, J., Koprivová, J., Novák, T., & Adamcová, K. (2009). Physiological dissociations in obsessive-compulsive disorder is associated with anxiety levels but not with severity of obsessive-compulsive symptoms. *Neuroendocrinology Letters*, 30, 625-628.
- Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, 11, 306-322.
- Rosario-Campos, M. C., Miguel, E. C., Quatrano, S., Chacon, P., Ferrao, Y., Findley, D., . . . Leckman, J. F. (2006). The Dimensional Yale-Brown Obsessive-Compulsive Scale (DY-BOCS): An instrument for assessing obsessive-compulsive symptom dimensions. *Molecular Psychiatry*, 11, 495-504.
- Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *British Journal of Mathematical and Statistical Psychology*, 66, 402-425.
- Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, 13, 31-48.
- Sawilowski, S. (2000). Psychometrics versus datametrics: comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Educational and Psychological Measurement*, 60, 157-173.
- Scahill, L., Riddle, M. A., McSwiggin-Hardin, M., Sharon, I. O., King, R. A., Goodman, W. K., . . . Leckman, J. F. (1997). Children's Yale-Brown Obsessive-Compulsive Scale: Reliability and Validity. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36, 844-852.
- Shields, A. L., & Caruso, J. C. (2004). A reliability induction and reliability generalization study of the Cage Questionnaire. *Educational and Psychological Measurement*, 64, 254-270.
- Sica, C., Coradeschi, D., Sanavio, E., Dorz, S., Manchisi, D., & Novara, C. (2004). A study of the psychometric properties of the Obsessive Beliefs Inventory and interpretations of Intrusions Inventory on clinical Italian individuals. *Journal of Anxiety Disorders*, 18, 291-307.
- Solem, S., Hjemdal, O., Vogel, P. A., & Stiles, T. C. (2010). A Norwegian version of the Obsessive-Compulsive Inventory-Revised: Psychometric properties. *Scandinavian Journal of Psychology*, 51, 509-516.
- Somers, J. M., Goldner, E. M., Waraich, P., & Hsu, L. (2006). Prevalence and incidence studies of anxiety disorders: A systematic review of the literature. *Canadian Journal of Psychiatry*, 51, 100-113.
- Storch, E. A., Shapira, N. A., Dimoulas, E., Geffken, G. R., Murphy, T. K., & Goodman, W. K. (2005). Yale-Brown obsessive compulsive scale: The dimensional structure revisited. *Depression and Anxiety*, 22, 28-35.
- Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). New York, NY: Oxford University Press.
- Summerfeldt, L. J., Richter, M. A., Antony, M. M., & Swinson, R. P. (1999). Symptom structure in obsessive-compulsive disorder: A confirmatory factor-analytic study. *Behaviour Research and Therapy*, 37, 297-311.
- Taylor, S. (2011). Early versus late onset obsessive-compulsive disorder: Evidence for distinct subtypes. *Clinical Psychology Review*, 31, 1083-1100.
- Tek, C., Ulug, B., Gürsoy-Rezaki, B., Tanriverdi, N., Mercan, S., Demir, B., & Vargel, S. (1995). Yale-Brown Obsessive Compulsive Scale and US National Institute of Mental Health Global Obsessive Compulsive Scale in Turkish: Reliability and validity. *Acta Psychiatrica Scandinavica*, 91, 410-413.
- Thompson, B. (Ed.). (2003). *Score reliability*. Thousand Oaks, CA: Sage.
- 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., & 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.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. Retrieved from <http://www.jstatsoft.org/v36/i03/paper>
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.