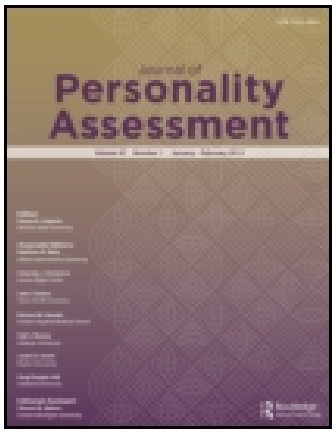

Journal of Personality Assessment

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hjpa20

Reliability Generalization Study of the Yale-Brown Obsessive-Compulsive Scale for Children and Adolescents

José Antonio López-Pina[a], Julio Sánchez-Meca[a], José Antonio López-López[a], Fulgencio Marín-Martínez[a], Rosa Ma Núñez-Núñez[b], Ana I. Rosa-Alcázar[c], Antonia Gómez-Conesa[d] & Josefa Ferrer-Requena[a]

[a] Department of Basic Psychology and Methodology, University of Murcia, Spain

[b] Department of Health Psychology, Miguel Hernández University at Elche, Spain

[c] Department of Personality, Assessment, and Psychological Treatment, University of Murcia, Spain

[d] Department of Physical Therapy, University of Murcia, Spain

Published online: 10 Jul 2014.

To cite this article: José Antonio López-Pina, Julio Sánchez-Meca, José Antonio López-López, Fulgencio Marín-Martínez, Rosa Ma Núñez-Núñez, Ana I. Rosa-Alcázar, Antonia Gómez-Conesa & Josefa Ferrer-Requena (2015) Reliability Generalization Study of the Yale-Brown Obsessive-Compulsive Scale for Children and Adolescents, Journal of Personality Assessment, 97:1, 42-54, DOI: 10.1080/00223891.2014.930470

To link to this article: http://dx.doi.org/10.1080/00223891.2014.930470

Journal of Personality Assessment, *97*(1), 42–54, 2015

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223891.2014.930470

Routledge
Taylor & Francis Group

Reliability Generalization Study of the Yale–Brown Obsessive–Compulsive Scale for Children and Adolescents

José Antonio López-Pina,[1] Julio Sánchez-Meca,[1] José Antonio López-López,[1] Fulgencio Marín-Martínez,[1] Rosa Mª Núñez-Núñez,[2] Ana I. Rosa-Alcázar,[3] Antonia Gómez-Conesa,[4] and Josefa Ferrer-Requena[1]

[1]*Department of Basic Psychology and Methodology, University of Murcia, Spain*
[2]*Department of Health Psychology, Miguel Hernández University at Elche, Spain*
[3]*Department of Personality, Assessment, and Psychological Treatment, University of Murcia, Spain*
[4]*Department of Physical Therapy, University of Murcia, Spain*

The Yale–Brown Obsessive–Compulsive Scale for children and adolescents (CY–BOCS) is a frequently applied test to assess obsessive–compulsive symptoms. We conducted a reliability generalization meta-analysis on the CY–BOCS to estimate the average reliability, search for reliability moderators, and propose a predictive model that researchers and clinicians can use to estimate the expected reliability of the CY–BOCS scores. A total of 47 studies reporting a reliability coefficient with the data at hand were included in the meta-analysis. The results showed good reliability and a large variability associated to the standard deviation of total scores and sample size.

According to the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed. [*DSM–5*]; American Psychiatric Association, 2013) 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 experienced 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, whereas compulsions are repetitive and stereotyped behaviors that require specific rules to be followed, their goal being to reduce the distress caused by obsessions. Persons who suffer from this disorder can attract the attention of other people due to their irrational and unadapted behavior. In addition, the disorder could lead to great functional impairment and disability in the person (Abramowitz, Taylor, & McKay, 2009; Taylor, 2011). OCD is suffered not only by adults, but also by children and adolescents, and its consequences in their lives can be severe. OCD causes considerable interference in children's activities (play, academic tasks, social life, etc.) due to the enormous amount of time lost to compulsions. Prevalence rates reported by some studies in the child and adolescent population have reached figures around 2% to 4% (Apter et al., 1996; Maina, Albert, Bogetto, & Ravizza, 1999; Rapoport et al., 2000; Zohar, 1999).

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 nonclinical populations for screening purposes (Storch, Benito, & Goodman, 2011).

Received April 30, 2013; Revised May 2, 2014.

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

Although the Y–BOCS was initially developed to assess adults with OCD, it has been adapted for children and adolescents (CY–BOCS) as well as other psychiatric disorders in which obsessions and compulsions play a relevant role. The most frequently used adaptations are those for body dysmorphic disorder (Phillips et al., 1997), eating disorders (Mazure, Halmi, Sunday, Romano, & Einhorn, 1994), pathological gambling (Hollander et al., 1998), heavy drinking (Modell, Glaser, Mountz, Schmaltz, & Cyr, 1992), and compulsive shopping (Monahan, Black, & Gabel, 1996).

The focus of this investigation was the CY–BOCS, and we were interested both in the original version and in the test adaptations to various languages and cultures (Alaghband-Rad & Hakimshooshtary, 2009; Alvarenga et al., 2006; Ben-Dor et al., 2007; Brynska & Wolanczyk, 2005; Chabane et al., 2005; Cubo et al., 2008; Guldeniz-Yucelen, Rodopman-Arman, Topcuoglu, Yanki-Yazgan, & Fisek, 2006; Ivarsson & Valderhaug, 2006; Ivarsson & Winge-Westholm, 2004; Kim, Yoo, Soo-Churcl, Kang-E, & Boong-Nyun, 2005; Steinberg & Schuch, 2002; Termine et al., 2006; Thomsen, 1994; Verhaak & De Haan, 2007; Wang & Kuo, 2003).

The CY–BOCS maintains the same structure as that of the original version of the Y–BOCS for adults: a 10-item, clinician-rated, semistructured instrument that assesses the presence and severity of obsessions and compulsions over the last week. Five items are intended to assess the severity of obsessions, and the remaining five address the assessment of compulsions. All items have a Likert-type scale scored from 0 (*none*) to 4 (*extreme*), so that the test provides a total score by adding the 10 items (range = 0–40), as well as specific scores for the Obsessions and Compulsions subscales (range = 0–20). In addition, the CY–BOCS assesses the presence of a list of items concerning insight, avoidance, indecisiveness, pathological doubting, obsessive slowness, and overvalued ideation. The focus of this investigation is on the reliability of total scores, as well as that of the Obsessions and Compulsions scores.

An important aspect of any measurement instrument is the psychometric quality of its scores. Reliability is one of the

most relevant properties of the test scores. It can be defined as the consistency of measurement over the testing conditions (Anastasi & Urbina, 1997). Such conditions include content sampling (e.g., coefficient alpha), and interrater differences (e.g., intraclass correlation). An appropriate reliability of the test scores is crucial for the inclusion of these data in the diagnostic process. Moreover, 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 an assessment instrument, reliability is a prerequisite toward achieving valid conclusions in both clinical and research contexts (Nunnally & Bernstein, 1994).

In the original study, Scahill et al. (1997) applied the CY–BOCS to a group of 65 children with OCD, finding an internal consistency of .87 and intraclass correlations of .84, .91, and .66 for CY–BOCS Total, Obsessions, and Compulsions scores, respectively. To our knowledge, 10 psychometric studies of the CY–BOCS have been published (Adrianzen-Ronceros, Pacheco-Armas, Vivar-Cuba, & Macciota-Felices, 2008; Freeman, Flessner, & García, 2011; Godoy et al., 2011; Guldeniz-Yucelen et al., 2006; McKay et al., 2003; Scahill et al., 1997; Storch, Murphy, Adkins, et al., 2006; Storch, Murphy, Geffken, Bagner, et al., 2005; Storch et al., 2004a, 2004b; Ulloa et al., 2004). Their results offer good internal consistency (coefficients alpha between .66–.90) and interrater agreement (intraclass correlations between .79–.94) overall. Nonetheless, these studies also show a clear heterogeneity in reliability estimates depending on sample composition and variability. In addition, it is not clear whether the large number of different adaptations of the CY–BOCS to other languages and cultures exhibit similar reliability estimates in 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 induce score reliability from previous administrations of the test to other samples. Reliability induction is a problematic practice because, as psychometric theory states, reliability is not a property of the test itself, but of the scores obtained in a given administration of the test to a sample of participants and in a given context (Crocker & Algina, 1986; Lord & Novick, 1968; McDonald, 1999; Streiner & Norman, 2008). Therefore, researchers should compute a reliability coefficient using the test scores obtained from the persons under assessment.

As score reliability changes from one 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. In this respect, meta-analysis constitutes a suitable method that allows the examination of how score reliability varies throughout 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 study, an exhaustive search of the studies that have applied the test is carried out and those that report any reliability estimate based on the study-specific sample are included in the meta-analysis (Henson & Thompson, 2002; Rodriguez & Maeda, 2006; Rouse, 2007; Sánchez-Meca, López-López, & López-Pina, 2013; Vacha-Haase & Thompson, 2011). Since 1998, more than 80 RG meta-analyses have been published or carried out, such as those of Rouse (2007) on the Minnesota Multiphasic Personality Inventory–2 PSY–5 scales, Vassar and Bradley (2010) on the Life Orientation Test, or Vassar and Crosby (2008) on the UCLA Loneliness Scale. The wide use of the CY–BOCS justifies the need for analyzing how CY–BOCS scores vary throughout different test administrations.

Purpose

We conducted an RG study on the CY–BOCS to (a) estimate the average reliability, in terms of internal consistency and interrater agreement, obtained in the empirical studies that applied the CY–BOCS; (b) examine variability among the reliability estimates; (c) search for substantive and methodological characteristics of the studies that can be statistically associated to test score reliability coefficients, if there is more variability than sampling error can explain; and (d) propose a predictive model that researchers and clinicians can use in the future to estimate the expected reliability of the CY–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 standard deviation of the test scores, mean examinee age, and test version (original vs. adapted) would all affect the score reliability.

Method

Data Sources

Although the CY–BOCS was developed in 1997, it was adapted from the original Y–BOCS for adults from 1989, so that the search period of the relevant studies covered 1989 to 2011 inclusive. The following databases were consulted: MedLine, SCOPUS, PsycInfo, PUBMED, and PROQUEST CENTRAL. In all of the electronic databases, the following keywords were combined to be found not only in the title or the abstract, but throughout the document: *Yale–Brown Obsessive Compulsive Scale, Y–BOCS, YBOCS,* or *YBOC.* A complementary electronic search was launched combining these keywords: *factor analysis* and *Y–BOCS* or *reliability* and *Y–BOCS* or *validity* and *Y–BOCS.* By using keywords, in contrast to subject headings specific to particular databases, we intended to create a broad search strategy and avoid unintentional exclusion of any study.

Study Selection

To be included in the meta-analysis, the study had to fulfill three criteria: (a) to be a study where the CY–BOCS was applied to a sample of children, adolescents, or both; (b) to report any reliability estimate based on the study-specific sample; and (c) due to language limitations, the study had to be written in English, Spanish, or French.

The search yielded a total of 11,490 references, out of which 11,145 were removed for different reasons (7,067 were duplicates, 1,560 were not empirical studies, and 2,518 were studies that did not apply the CY–BOCS, but either the original Y–BOCS or some other adaptation of the latter). The remaining 345 references were studies that had applied the CY–BOCS for children. Out of these, 47 (13.6%) studies reported any estimate of the test score's reliability, whereas the remaining 298 (86.4%) induced reliability from other

studies. Two kinds of reliability induction can be distinguished when researchers do not report a reliability estimate of test scores with the data at hand (Shields & Caruso, 2004): Reliability induction by omission consists of omitting any reference to the score reliability, whereas reliability induction by report occurs when the study reports a reliability estimate from previous studies. Out of the 298 studies that induced reliability, 220 (63.8%) omitted any reference to the reliability of the CY–BOCS, whereas the remaining 78 studies (22.6%) induced reliability by reporting a previous reliability estimate.

Data Extraction

To explore how study characteristics can affect score reliability when the CY–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) mean age of participants (in years), (e) standard deviation of the age of participants (in years), (f) gender distribution in the sample (% male), (g) target population of the sample (nonclinical vs. clinical), (h) disorder of the participants (OCD vs. other), (i) study focus (psychometric vs. substantive), (j) focus of the psychometric studies (Y–BOCS vs. other tests), (k) publication year, (l) country where the study was conducted (United States vs. other), (m) discipline of main researcher (psychology vs. psychiatry), and (n) sample size. As reliability estimates are sample-specific, some selected moderators were chosen to reflect sample characteristics and heterogeneity (e.g., mean age, standard deviation of age, gender distribution, target population). Moreover, some methodological and extrinsic characteristics were considered to examine their relationship with the reliability estimates. The discipline of the main researcher was recorded, as other RG studies have also done so, based on the assumption that there can be differences in the reporting practices of reliability between psychological and medical fields (see, e.g., Barnes, Harp, & Jung, 2002).

Together with these moderator variables, coefficients alpha and intraclass correlations, as well as other types of reliability estimates, were obtained for the total scale and for the subscales when these were reported in the studies. Of the 47 studies that did not induce reliability, 46 reported coefficients alpha, intraclass correlations, or kappa coefficients, whereas one study provided a Spearman's rho coefficient (Dewrang & Sandberg, 2011). Therefore, the database of our RG study was based on the 46 studies that reported any coefficient alpha, intraclass correlation, or kappa coefficient (see Appendix).

To check the reliability of the coding process of the study characteristics, all studies included in the meta-analysis were doubly coded by two independent coders. The results were highly satisfactory overall, with kappa coefficients ranging between .74 and 1.0 for the qualitative characteristics, and intraclass correlations ranging between .78 and 1.0 for the continuous variables (Orwin, 1994). The inconsistencies between the coders were solved by consensus.

Data Analysis

Separate meta-analyses were conducted for each type of reliability coefficients, as they estimate different types of reliability, following the recommendations of several authors from the RG arena (e.g., Henson & Thompson, 2002). There is some debate about whether or not reliability coefficients should be transformed for their statistical integration. Some authors recommend not transforming reliability coefficients aiming to ease the interpretation of the meta-analytic results (Henson & Thompson, 2002; Leach, Henson, Odom, & Cagle, 2006; Mason, Allam, & Brannick, 2007; Thompson & Vacha-Haase, 2000). Other authors propose transforming reliability coefficients to normalize their distribution and stabilizing their sampling variances (Feldt & Charter, 2006; Rodriguez & Maeda, 2006; Sánchez-Meca et al., 2013). As a sensitivity analysis, we carried out the statistical analyses both with untransformed and transformed reliability coefficients. Thus, coefficients alpha were transformed by means of Bonett's (2002) formula: $T_i = Ln(1 - |\hat{\alpha}_i|)$, Ln being the natural logarithm, T_i being the transformed coefficient, and $\hat{\alpha}_i$ being the coefficient alpha for the ith study. In addition, intraclass correlations were transformed using Fisher's Z: $Z_i = 0.5\ Ln[(1+\hat{\alpha}_i)/(1-\hat{\alpha}_i)]$ (Sánchez-Meca et al., 2013).

To obtain summary statistics of reliability coefficients, a random-effects model was 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). The sampling variances of untransformed coefficients alpha were estimated by means of (Bonett, 2010; Sánchez-Meca et al., 2013):

$$V(\hat{\alpha}_i) = \frac{2J(1-\hat{\alpha}_i)^2}{(J-1)\left\{N_i - 2 - [(J-2)(k-1)]^{1/4}\right\}},$$

with J being the number of items of the test, k being the number of studies in the meta-analysis, and N_i being the sample size of the ith study. For transformed coefficients alpha, the sampling variances were obtained by:

$$V(T_i) = \frac{2J}{(J-1)(N_i-2)}.$$

The sampling variances of the untransformed intraclass correlations, r_i, were estimated by means of:

$$V(r_i) = \frac{(1-r_i^2)^2}{(N_i-2)}.$$

And for the intraclass correlations transformed by the Fisher's Z, the sampling variances were obtained by:

$$V(Z_i) = \frac{1}{(N_i-3)}.$$

In all cases, the confidence limits around the overall reliability estimate were computed using the method proposed by Hartung, which assumes a Student t distribution with $k - 1$ degrees of freedom (k being the number of reliability coefficients) and an improved estimate of the sampling variance of the mean reliability coefficient (cf. Hartung, 1999; Sánchez-Meca & Marín-Martínez, 2008). The heterogeneity exhibited

by the reliability estimates was assessed with the Q statistic and the I^2 index (Borenstein, Hedges, Higgins, & Rothstein, 2009). The Q statistic can be applied to test the homogeneity assumption among the reliability coefficients, and the I^2 index allows expression of the amount of heterogeneity as a percentage (Higgins & Thompson, 2002). I^2 values around 25%, 50%, and 75% can be considered as reflecting moderate, substantial, and considerable heterogeneity, respectively (Deeks, Higgins, & Altman, 2008).

If studies were found to have heterogeneity, then moderator analyses were conducted through regression analyses for the continuous variables and analyses of variance (ANOVAs) for the qualitative ones. Mixed-effects models were assumed for these analyses, using the adjustment proposed by Knapp and Hartung to test the statistical significance of the moderator variable (Knapp & Hartung, 2003; López-López, Botella, Sánchez-Meca, & Marín-Martínez, 2013; Sánchez-Meca et al., 2013). The t and Q_B statistics for regression analyses and ANOVAs, respectively, allow testing whether a moderator variable is statistically associated to the reliability coefficients, whereas the Q_E and Q_W statistics, respectively, enable us to examine the model misspecification. The proportion of variance accounted for by the moderator variable was estimated by means of $R^2_{adj} = 1 - \hat{\tau}^2_{Res}/\hat{\tau}^2$, $\hat{\tau}^2_{Res}$ and $\hat{\tau}^2$ being the residual and total between-studies variances, respectively (López-López et al., 2014; Raudenbush, 2009).

To facilitate interpretation of results, the average reliability estimates and their confidence limits 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).

Finally, some sensitivity analyses were conducted. As noted earlier, for comparison purposes the statistical analyses were carried out both with the untransformed and transformed reliability coefficients. Moreover, the risk of publication bias was assessed constructing funnel plots and applying the trim-and-fill method (Duval & Tweedie, 2000). Both methods are intended to check if the review failed to include studies with reliability estimates lower than desirable. A funnel plot allows this threat to be assessed graphically, whereas the number of missing reliability coefficients can be estimated with the trim-and-fill method.

RESULTS

Descriptive Characteristics of the Studies

This RG study focused on the 46 studies that reported any coefficient alpha, intraclass correlation, or kappa coefficient. Most studies (43) were written in English, and the three remaining articles were written in Spanish. Thirty-eight (82.6%) studies were conducted in the United States, and the remaining eight studies were carried out in different parts of Europe, Asia, and South America. In our pooled sample, ages ranged between 3 and 18 years, with an overall mean age of 12.6 years. On average, each sample included 52.9% males and the mean values for the average total CY–BOCS score and its standard deviation were 16.6 and 5.9, respectively.

Of the 46 studies, 45 applied the CY–BOCS in a clinician-rated format and only one applied it as a self-report (Godoy et al., 2011). This implies that the CY–BOCS scores obtained from the clinician-rated format were assigned by a clinician taking into account the answers of children and their parents to the items. On the contrary, in the Godoy et al. study, the CY–BOCS scores were obtained directly from the children ratings.

The most frequently reported reliability estimate for the total scores was coefficient alpha, computed from 33 (71.7%) studies and leading to a pooled sample of $N = 3{,}663$ participants. Other types of reliability found were interrater agreement coefficients, with the intraclass correlation reported in 18 (39.1%) studies, leading to a total sample of $N = 1{,}113$ participants, and the kappa coefficient computed for 9 (19.6%) samples, for a total of $N = 662$ persons. Regarding the Obsessions and Compulsions subscales, coefficient alpha was reported in 8 (17.4%) studies from a total of $N = 2{,}314$ participants, and the intraclass correlation was computed for 6 (13%) samples, using a total of $N = 405$ persons. Separate meta-analyses were conducted for coefficients alpha and intraclass correlations. Kappa coefficients were not analyzed due to the small number of studies reporting them.

A detailed inspection of the data revealed a study (Godoy et al., 2011) with a coefficient alpha of .58, clearly out of the range of .71 to .95 exhibited by the remaining studies. The intraclass correlation reported in this study (.66) was also out of the range (.79–.99) of the studies. In addition, its sample size ($N = 1{,}706$) was the largest (range of the remaining studies: 7–500). Both for the coefficients alpha and the intraclass correlations, this study added a large heterogeneity and it was not representative of the remaining integrated studies. Therefore, as a sensitivity analysis all of the statistical analyses were doubly carried out including and removing this study.

Mean Reliability and Heterogeneity

Table 1 shows the main summary statistics for coefficients alpha with the 33 studies and once deleting the Godoy et al. (2011) study. With comparison purposes, both the results with raw (α) and transformed (T) coefficients were presented.

TABLE 1.—Reliability and 95% confidence intervals for the raw (α) and transformed (T) coefficients alpha for the total scale and the Obsessions and Compulsions subscales.

	k	Min.	Max.	M	95% CI	Q	I^2
Total scale (α)	33	.58	.95	.84	[.81, .87]	642.22**	95.02
Total scale (T)	33	.58	.95	.85	[.82, .87]	788.74**	95.94
Total scale (α)[a]	32	.71	.95	.86	[.84, .88]	181.04**	82.88
Total scale (T)[a]	32	.71	.95	.86	[.83, .88]	218.18**	85.79
Obsessions (α)	8	.32	.92	.76	[.60, .93]	491.14**	98.57
Obsessions (T)	8	.32	.92	.81	[.66, .90]	561.96**	98.75
Obsessions (α)[a]	7	.64	.92	.85	[.78, .92]	26.99**	77.77
Obsessions (T)[a]	7	.64	.92	.85	[.76, .91]	45.88**	86.92
Compulsions (α)	8	.37	.94	.75	[.60, .89]	557.07**	98.74
Compulsions (T)	8	.37	.94	.79	[.63, .88]	535.07**	98.69
Compulsions (α)[a]	7	.71	.94	.82	[.75, .89]	61.58**	90.26
Compulsions (T)[a]	7	.71	.94	.82	[.71, .89]	90.35**	93.36

Note. To facilitate the interpretation, all means and their respective confidential limits were back-transformed to the metric of the original coefficients when some transformation was applied. k = number of studies (or reliability coefficients); Min. and Max. = minimum and maximum reliability coefficients, respectively.

[a]Results after removing the Godoy et al. (2011) study from the analyses.

**$p < .001$.

For the untransformed coefficients, the (weighted) mean coefficient alpha for the total scale was .84. The results were very similar when using the transformed coefficients, with an average of .85 and a slightly smaller confidence interval. When the Godoy et al. (2011) study was deleted from the analyses, the mean coefficient alpha was slightly larger than the one obtained for all of the studies ($M = .86$, both for the transformed and the untransformed coefficients). For the subscales, the analyses conducted with the raw coefficients alpha yielded an overall estimate of .76 for the Obsessions subscale and .75 for the Compulsions subscale. The overall reliability estimates were slightly larger when the transformed coefficients were integrated (Obsessions subscale = .81; Compulsions subscale = .79). When the Godoy et al. study was removed from the analyses, the mean coefficient alpha was slightly larger than when it was included (Obsessions subscale = .85; Compulsions subscale = .82). Table 1 also presents the results of the Q statistics and the I^2 indexes assessing the variability exhibited by the reliability estimates. Coefficients alpha for the total scale and subscales showed a statistically significant heterogeneity, with almost all I^2 values exceeding 80%. After removing the Godoy et al. study, the heterogeneity Q statistics reduced drastically, although still exhibiting a large variability. Consequently, analyses to explain part of that heterogeneity were needed.

Table 2 presents the main descriptive statistics for intraclass correlations. Similar to Table 1, both the results with the raw (r) and the transformed (Z) intraclass correlations are shown, with two rows for each to display values obtained with and without the Godoy et al. (2011) study, respectively.

Intraclass correlations for the total scale yielded an average reliability of .87 with the raw correlations and a slightly higher mean of .89 with the transformed estimates. When the Godoy et al. (2011) study was removed from the analyses, the mean coefficient was slightly larger (means of .89 and .90 for the untransformed and transformed coefficients, respectively). For the Obsessions subscale, a mean estimate of .82 was obtained and a mean of .76 was obtained for the Compulsions subscale. The overall estimates for both subscales were very similar when using transformed correlations (.83 and .75 for Obsessions and Compulsions, respectively), although these integrations produced wider confidence intervals. Slightly larger average estimates were found when the Godoy et al. study was removed, with values of .84 and .81 for the Obsessions and Compulsions subscales, respectively, for the untransformed correlations, and values of .85 and .79 for the transformed ones. Finally, the results of the Q statistics and the I^2 indexes indicated the presence of heterogeneity, although with a clear reduction when the Godoy et al. study was removed from the analyses.

TABLE 2.—Reliability and 95% confidence intervals for the raw (r) and transformed (Z) intraclass correlations for the total scale and the Obsessions and Compulsions subscales.

	k	Min.	Max.	M	95% CI	Q	I^2
Total scale (r)	18	.66	.99	.87	[.83 .91]	238.57**	92.87
Total scale (Z)	18	.66	.99	.89	[.83 .93]	159.67**	89.35
Total scale (r)[a]	17	.79	.99	.89	[.85 .92]	159.58**	89.97
Total scale (Z)[a]	17	.79	.99	.90	[.84 .93]	90.88**	82.39
Obsessions (r)	6	.66	.94	.82	[.68 .95]	46.94**	89.35
Obsessions (Z)	6	.66	.94	.83	[.63 .92]	26.51**	81.14
Obsessions (r)[a]	5	.66	.94	.84	[.69 1.00]	14.74*	72.86
Obsessions (Z)[a]	5	.66	.94	.85	[.62 .95]	18.64**	78.54
Compulsions (r)	6	.61	.89	.76	[.63 .88]	33.02**	84.86
Compulsions (Z)	6	.61	.89	.75	[.59 .86]	21.32**	76.55
Compulsions (r)[a]	5	.66	.89	.81	[.70 .93]	7.35	45.58
Compulsions (Z)[a]	5	.66	.89	.79	[.63 .89]	8.03	50.19

Note. To facilitate the interpretation, all means and their respective confidential limits were back-transformed to the metric of the original coefficients when some transformation was applied. k = number of studies (or reliability coefficients); Min. and Max. = minimum and maximum reliability coefficients, respectively.

[a]Results after removing the Godoy et al. (2011) study from the analyses.

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

Moderator Analyses

Tables 3 and 4 present the results of the moderator analyses for quantitative and categorical variables, respectively, on coefficients alpha. Only results using the transformed coefficients are discussed here, although the moderator analyses with the raw coefficients alpha were also conducted and the results were comparable (tables with the results for the untransformed coefficients alpha can be obtained from the corresponding author on request). All the moderator analyses were carried out both including and excluding the Godoy et al. (2011) study. When the exclusion of this study led to relevant changes in the results of a moderator variable, it was made explicit.

Table 3 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. Note that the sign of the regression slopes, b_j, reported in Table 3 are those obtained taking as the dependent variable the coefficients alpha transformed by Bonett's (2002) formula. This means that the true relationship between the raw coefficients alpha and a moderator is the inverse of that represented by the sign of the slope in Table 3. For example, the negative relationship found for the standard deviation of the total scores with transformed coefficients alpha would actually be interpreted as a positive relationship between raw coefficients alpha and the moderator variable, and vice versa. As psychometric theory predicts, a positive, statistically significant relationship was found between the standard deviation of the total test scores and the reliability estimates, with 35% of variance accounted for. The mean of the total scores did not show a statistical relationship with the coefficients alpha. However, when the Godoy et al. (2011) study was removed, a

TABLE 3.—Results of the simple metaregression 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	26	−0.113	−3.31	.003	.35	267.36**
Mean of the total scores	27	−0.012	−0.53	.599	.00	137.21**
Sample size	33	0.001	2.29	.029	.14	246.21**
Mean age (in years)	32	0.002	0.03	.974	.00	587.30**
SD of the age (in years)	29	−0.360	−2.13	.042	.13	207.87**
Percentage of males in the sample	32	0.002	0.20	.843	.00	724.13**
Year of publication	33	0.074	3.15	.004	.25	342.15**

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.

**$p < .001$.

TABLE 4.—Results of the weighted analyses of variance assuming a mixed-effects model on the transformed alpha coefficients for the categorical moderator variables.

Moderator Variable	k_j	$\overline{\alpha}_j$	95% CI	ANOVA Results
Test version				$Q_B = 0.10, p = .756$
Original	26	0.85	[0.82, 0.88]	$R^2_{adj} = 0$
Adapted	7	0.84	[0.77, 0.89]	$Q_W = 432.93, p < .001$
Study focus				$Q_B = 0.27, p = .607$
Psychometric	14	0.86	[0.82, 0.89]	$R^2_{adj} = 0$
Substantive	19	0.85	[0.80, 0.88]	$Q_W = 679.34, p < .001$
Psychometric focus				$Q_B = 0.80, p = .390$
CY–BOCS	10	0.85	[0.77, 0.90]	$R^2_{adj} = 0$
Other	4	0.89	[0.79, 0.94]	$Q_W = 526.40, p < .001$
Country				$Q_B = 0.10, p = .756$
United States	26	0.85	[0.82, 0.88]	$R^2_{adj} = 0$
Other	7	0.84	[0.77, 0.89]	$Q_W = 432.93, p < .001$
Target population				$Q_B = 0.97, p = .333$
Nonclinical	2	0.80	[0.62, 0.89]	$R^2_{adj} = 0$
Clinical	31	0.85	[0.83, 0.88]	$Q_W = 352.99, p < .001$
Disorder				$Q_B = 1.57, p = .219$
OCD	28	0.85	[0.82, 0.87]	$R^2_{adj} = .02$
Other	3	0.89	[0.82, 0.94]	$Q_W = 208.58, p < .001$
Diagnosis				$Q_B = 0.27, p = .844$
DSM–III–R	1	0.87	[0.67, 0.95]	$R^2_{adj} = 0$
DSM–IV	13	0.86	[0.81, 0.89]	
DSM–IV–TR	9	0.84	[0.79, 0.89]	$Q_W = 203.14, p < .001$
Other	9	0.87	[0.82, 0.91]	
Researcher affiliation				$Q_B = 1.30, p = .263$
Psychologist	8	0.82	[0.75, 0.88]	$R^2_{adj} = .01$
Psychiatrist	23	0.86	[0.83, 0.89]	$Q_W = 519.91, p < .001$

Note. To facilitate the interpretation, the average reliability coefficients and their respective confidence limits were back-transformed to the metric of the original coefficients. k_j = number of studies (or coefficients) for each category of the moderator variable; $\overline{\alpha}_j$ = average reliability coefficient for each category of the moderator variable; ANOVA = analysis of variance; 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; CY–BOCS = Yale–Brown Obsessive–Compulsive Scale for children and adolescents; OCD = obsessive–compulsive disorder; *DSM–III–R = Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.); *DSM–IV = Diagnostic and Statistical Manual of Mental Disorders* (4th ed.); *DSM–IV–TR = Diagnostic and Statistical Manual of Mental Disorders* (4th ed, Text revision).

negative, marginally significant result was found, $t(24) = 2.02$, $p = .054$, $R_{adj}^2 = .16$. In addition, a negative, statistically significant relationship was found between the sample sizes and the reliability estimates, with 13.7% of variance accounted for. When the Godoy et al. study was removed, a statistically significant relationship was also found with the coefficients alpha, $t(30) = -2.60$, $p = .014$, $R_{adj}^2 = .20$, but in this case the direction of the relationship reversed from negative to positive. The standard deviation of the age showed a positive, statistically significant relationship with coefficients alpha. However, when the Godoy et al. study was removed, the statistical relationship disappeared, $t(26) = -1.25$, $p = .223$, $R_{adj}^2 = .03$. The study year showed a negative relationship with the coefficients, with 25.4% of variance explained. The mean age and gender distribution (% males) did not reach the statistical significance.

With regard to the qualitative moderators, Table 4 shows the results of the weighted ANOVAs applied on the transformed coefficients alpha of the total scale. None of the moderators included in this table yielded a statistically significant relationship with the coefficients. However, there were three moderator variables with mean coefficients that reversed when the Godoy et al. (2011) study was removed from the analyses. Studies using adapted versions of the CY–BOCS showed a smaller average coefficient ($M = .84$) than those with original test applications ($M = .85$) when the Godoy et al. study was included in the analysis, but larger when this study was excluded ($M = .87$). In addition, studies conducted in the United States had a smaller mean coefficient than the one obtained for those carried out in the United States when the Godoy et al. study was included, but larger when it was excluded.

Similarly, Tables 5 and 6 show results of the moderator analyses for the same continuous and categorical variables, but now using the transformed intraclass correlations for the total scale as the dependent variable.

Table 5 presents the results of the weighted simple regression analyses conducted for the continuous moderators. None of the moderators was found to be statistically associated with the correlations, not even the standard deviation of the total scores, as the psychometric theory predicts. Marginally significant results were found for sample size and for the standard deviation of the age. However, when the Godoy et al. (2011) study was removed from the analyses, these moderators did not show any association with the reliability coefficients: sample size, $t(15) = -0.87$, $p = .397$; *SD* of age, $t(12) = 1.38$, $p = .193$. Regarding the qualitative moderators, Table 6 presents the results of the weighted ANOVAs applied on the transformed intraclass correlations for the total scale. The psychometric focus showed a statistically significant association with the correlations, with a higher reliability estimate for the only psychometric study that was not intended to analyze the CY–BOCS. Moreover, the diagnosis reference was also found to be statistically associated with the coefficients, with the highest intraclass correlations obtained when the DSM–IV and DSM–IV–R manuals were considered for the participants' diagnosis. The remaining moderator variables did not reach statistical significance (Table 6). However, it is worth noting that when the Godoy et al. study was deleted from the analyses, changes in the mean intraclass correlations were found. The mean coefficient was .89 for the original test administrations, whereas for adapted versions of the CY–BOCS the mean was .88 and .92 depending on whether the Godoy et al. (2011) study was included or deleted, respectively. The studies conducted in the United States showed a mean coefficient of .89, whereas those carried out in other countries changed their mean from .88 to .92 after removing the Godoy et al.

TABLE 5.—Results of the simple metaregression analyses assuming a mixed-effects model on the transformed intraclass correlation coefficients for the continuous moderator variables.

Moderator Variable	k	b_j	t	p	R_{adj}^2	Q_E
SD of the total scores	13	0.019	0.22	.832	.00	137.17**
Mean of the total scores	13	0.008	0.30	.768	.00	108.42**
Sample size	18	−0.003	−1.84	.086	.15	102.92**
Mean age (in years)	16	0.031	0.54	.598	.00	143.14**
SD of the age (in years)	15	0.450	1.99	.068	.20	89.92**
Percentage of males in the sample	16	0.005	0.43	.672	.00	121.19**
Year of publication	18	0.014	0.48	.640	.00	148.56**

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.

**$p < .001$.

TABLE 6.—Results of the weighted ANOVAs assuming a mixed-effects model on the transformed intraclass correlation coefficients for the categorical moderator variables.

Moderator Variable	k_j	$\overline{ICC}_j$	95% CI	ANOVA Results
Test version				$Q_B = 0.00, p = .955$
Original	13	0.89	[0.82, 0.93]	$R^2_{adj} = 0$
Adapted	5	0.88	[0.74, 0.95]	$Q_W = 125.47, p < .001$
Study focus				$Q_B = 0.00, p = .966$
Psychometric	7	0.89	[0.78, 0.94]	$R^2_{adj} = 0$
Substantive	11	0.89	[0.81, 0.94]	$Q_W = 138.08, p < .001$
Psychometric focus				$Q_B = 15.92, p = .010$
CY–BOCS	6	0.83	[0.68, 0.91]	$R^2_{adj} = .77$
Other	1	0.99	[0.94, 0.99]	$Q_W = 28.13, p < .001$
Country				$Q_B = 0.00, p = .955$
United States	13	0.89	[0.82, 0.93]	$R^2_{adj} = 0$
Other	5	0.88	[0.74, 0.95]	$Q_W = 125.47, p < .001$
Target population				$Q_B = 1.06, p = .319$
Nonclinical	2	0.81	[0.47, 0.94]	$R^2_{adj} = .004$
Clinical	16	0.90	[0.84, 0.93]	$Q_W = 136.68, p < .001$
Disorder				$Q_B = 0.17, p = .687$
OCD	13	0.89	[0.82, 0.93]	$R^2_{adj} = 0$
Other	4	0.91	[0.79, 0.96]	$Q_W = 87.01, p < .001$
Diagnosis				
DSM–III–R	1	0.84	[0.49, 0.96]	$Q_B = 6.38, p = .007$
DSM–IV	6	0.90	[0.84, 0.94]	$R^2_{adj} = .63$
DSM–IV–R	1	0.99	[0.96, 0.99]	$Q_W = 43.12, p < .001$
Other	9	0.86	[0.80, 0.90]	
Researcher affiliation				$Q_B = 0.52, p = .481$
Psychologist	7	0.87	[0.75, 0.94]	$R^2_{adj} = 0$
Psychiatrist	9	0.91	[0.83, 0.95]	$Q_W = 133.33, p < .001$

Note. To facilitate the interpretation, the average reliability coefficients and their respective confidence limits were back-transformed to the metric of the original coefficients. k_j = number of studies (or coefficients) for each category of the moderator variable; $\overline{ICC}_j$ = average reliability coefficient for each category of the moderator variable; ANOVA = analysis of variance; 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; CY–BOCS = Yale–Brown Obsessive–Compulsive Scale for children and adolescents; OCD = obsessive–compulsive disorder; *DSM–III–R = Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.); *DSM–IV = Diagnostic and Statistical Manual of Mental Disorders* (4th ed.); *DSM–IV–TR = Diagnostic and Statistical Manual of Mental Disorders* (4th ed, Text revision).

study. Studies with clinical samples obtained a mean coefficient of .90, whereas those that were nonclinical exhibited a mean coefficient of .81. However, when the Godoy et al. study was excluded, the differences between the two target populations disappeared. Finally, a similar result was found with the researcher affiliation. The mean intraclass correlations for psychologists and psychiatrists were .87 and .91, but these differences practically disappeared when the Godoy et al. study was removed from the analysis.

Although some moderators showed a statistically significant association with the reliability coefficients, none achieved a nonsignificant result for the model misspecification test (Q_E or Q_W for continuous and qualitative moderators, respectively). Therefore, a final objective of this meta-analysis was to propose an explanatory model containing the set of most relevant predictors.

An Explanatory Model

With the aim of finding a predictive model capable of explaining at least part of the variability among the reliability estimates, weighted multiple metaregression analyses were applied assuming a mixed-effects model. The small number of intraclass correlations ($k = 18$, with some factor categories including only one study, as shown in Table 6) led us to discard fitting a multiple regression model for this kind of reliability, so that we focused on the coefficients alpha for this purpose.

From a psychometric basis, it was expected that the standard deviation of test scores would be positively related with coefficients alpha. In addition to this moderator, another three moderator variables exhibited a statistically significant relationship with the coefficients alpha: the sample size, the standard deviation of the age, and the publication year of the study. To examine whether each of these three predictors reached statistical significance once the influence of the standard deviation of total scores was controlled, three multiple metaregressions were applied with two predictors each: the standard deviation of total scores and each of the other three predictors. To make the result interpretation easier, the raw coefficients alpha were taken as the dependent variable in the multiple metaregression models, instead of the transformed ones. In all three metaregressions, the standard deviation of test scores reached statistical significance, as did the sample size and the standard deviation of the age, but not the year of publication. Thus, in searching for the best explanatory model we conducted a metaregression with three predictors: the standard deviation of test scores, the sample size, and the standard deviation of the age. Although the full model was statistically significant, $F(3, 20) = 14.53$, $p < .0001$, $R_{adj}^2 = .71$, only the standard deviation of test scores, $t(20) = 2.57$, $p = .018$, and the sample size, $t(20) = -3.70$, $p = .001$, exhibited a statistically significant relationship with coefficients alpha, whereas the standard deviation of the age did not reach statistical significance, $t(20) = 1.04$, $p = .309$. Therefore, the standard deviation of the age was removed from the model, so that the best predictive model was that including the standard deviation of test scores and the sample size. The predictive model found was

$$\hat{\alpha}_i = 0.758 + 0.0138SD_i - 0.0001N_i.$$

The full model reached a statistically significant result, $F(2, 23) = 22.18$, $p < .001$, with 71.1% of variance accounted for. When individually testing the predictors, both the standard deviation of test scores, $t(23) = 2.75$, $p = .011$, and the sample size, $t(23) = -5.18$, $p < .001$, showed a highly statistically significant relationship with the coefficients alpha. As a counterpart, the model misspecification test was also statistically significant, $Q_E(23) = 57.60$, $p < .001$, therefore suggesting that other study characteristics were also affecting the coefficients alpha variability.

Publication Bias

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 1 presents the funnel plots obtained for coefficients alpha and intraclass correlations for the total scores. When the trim-and-fill method was applied on each funnel plot, no coefficients were imputed in the left side of the graph. Therefore, publication bias can be discarded as a threat to the meta-analytic results.

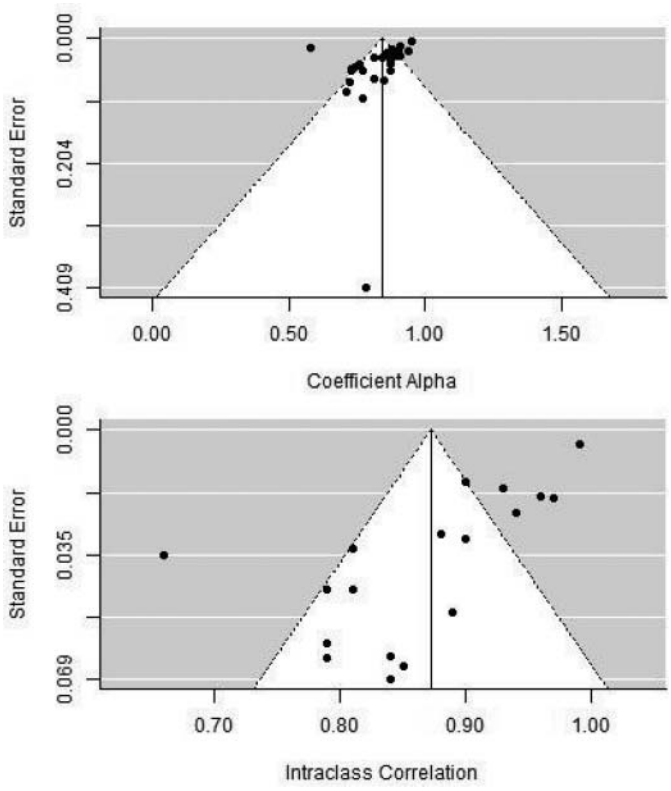

FIGURE 1.—Funnel plots of the reliability estimates for the total CY–BOCS scale.

DISCUSSION

As reliability is not a property of the test itself, but of the test scores obtained in each application, RG studies allow examining how reliability fluctuates through different test administrations, and to guide reliability expectations for future test applications. In this article, we presented the results of an RG study about the CY–BOCS, a commonly applied test for the assessment of obsessive–compulsive symptoms in children and adolescents.

The most commonly reported reliability estimate for the CY–BOCS was coefficient alpha, with a weighted average for the total scale of .85. The other type of reliability coefficients integrated were intraclass correlations, for which the overall reliability estimate for the total scale was approximately .89. Thus, on average both types of reliability were clearly over the cutoff of .70 usually considered as the minimum recommendable reliability when applying tests for exploratory research purposes, as well as when taking the limit of .80 for general research purposes (Nunnally & Bernstein, 1994). However, considering the more restrictive criterion of .90 for important clinical decisions, the average reliability did not comply with this criterion. The results of this RG study showed, therefore, that the CY–BOCS provides consistent information for its use with research purposes, but also that the scores should be interpreted cautiously when this instrument is applied in a clinical context.

The Godoy et al. (2011) study exhibited the lowest coefficient alpha ($\hat{\alpha}_i = .58$) and the largest sample size ($N = 1{,}706$). The main difference between this study and the rest was the administration format of the CY–BOCS: Whereas the Godoy et al. study used the self-report version of the CY–BOCS, the rest of the studies used a clinician-rated format. In addition, most of the studies of our meta-analysis applied the CY–BOCS to a clinical sample, whereas the Godoy et al. study applied it to a nonclinical sample. It is likely that children and adolescents have a lesser ability to identify obsessive–compulsive symptoms than trained clinicians (Gallant et al., 2008; Merlo, Storch, Murphy, Goodman, & Geffken, 2005). This circumstance can lead to a decrease in the variability of the CY–BOCS scores in the self-report version. A study carried out by Storch, Murphy, Adkins, et al. (2006) can shed light on this matter, as they applied the CY–BOCS, both as a clinician-rated format and as a self-report, to a sample of 53 children and adolescents with OCD. Coefficients alpha for the CY–BOCS clinical-rated and self-report versions were very similar, at .89 and .87, respectively. It is worth noting that the self-report version presented a lower average total score than the clinician-rated one and similar standard deviations (Ms = 14.2 and 19.9, SDs = 8.8 and 8.6, respectively). Thus, although the two versions exhibited similar coefficients alpha and standard deviations, the self-report version exhibited a lower mean total score than did the clinician-rated version. It is reasonable to assume that in a nonclinical sample, children and adolescents will have even more difficulties in identifying obsessive–compulsive symptoms when the CY–BOCS is applied as a self-report. These difficulties can lead to a diminished standard deviation of the CY–BOCS total scores and, as a consequence, a low coefficient alpha. This is what might have happened in the Godoy et al. (2011) study. Therefore, it does not seem advisable to apply the CY–BOCS in its self-report format, in particular to a nonclinical population.

Our results also showed a large variability among reliability estimates. Several characteristics of the studies presented a statistically significant relationship with both coefficients alpha and intraclass correlations. Results for the latter, however, should be tested in future RG studies, provided that the statistically significant associations were found for factors that included categories with a single value (see Table 6).

The moderator variable that exhibited the strongest relationship with coefficients alpha was the standard deviation of total scores. The influence of the score standard deviation is in agreement with psychometric theory, which predicts that the larger the score standard deviation, the larger the coefficient alpha (e.g., Crocker & Algina, 1986). Three other moderator variables exhibited a statistical relationship with coefficients alpha: the sample size, the year of publication, and the standard deviation of the age. However, once the influence of the standard deviation of total scores was controlled, only the sample size reached a statistical relationship with coefficients alpha. As a consequence, the predictive model proposed here only includes the standard deviation of total scores and the sample size as the two most relevant predictor variables of coefficient alpha.

These results have implications for researchers using the CY–BOCS. The predictive model proposed here can be used to make anticipations of the expected coefficient alpha as a function of the standard deviation of total scores and sample

size. For example, for the median standard deviation of total scores (median = 5.8) and the median sample size (median = 61) obtained in our RG study, the predicted coefficient alpha is .83.

It is also worth noting that the test version of the CY–BOCS revealed some differences in the reliability estimates. Both for coefficient alpha and intraclass correlation, administrations of the original CY–BOCS exhibited a slightly larger mean reliability than those of adapted versions. However, when the Godoy et al. (2011) study was excluded from the analyses, this trend was reversed. It is not clear, therefore, whether adapted versions of the CY–BOCS offer similar reliability in their scores.

Our results also have some clinical implications. On the one hand, the absence of a statistical relationship between coefficient alpha and personal characteristics of the samples, such as mean age of the sample, type of disorder, and gender distribution, suggests that the CY–BOCS yields appropriate reliability values regardless of the sample and administration conditions. On the other hand, our results present contradictory evidence regarding the coefficient alpha for clinical versus nonclinical samples, depending on the inclusion or not of the Godoy et al. (2011) study. This result points to the importance of reporting sample-specific reliability estimates so that the true trend can be explored in future research.

Although this RG study allowed us to identify several characteristics of the studies statistically related with the scores reliability of the CY–BOCS, there is also clear evidence of residual variability that remains to be explained, probably by other moderator variables not coded in our study. Another difficulty of our RG study was the presence of categories of the moderator variables with only a reliability estimate, as under these conditions the results of the statistical analyses are very unstable.

It is also worth noting that only 13.6% of the studies that applied the CY–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 in the scientific report. Therefore, our conclusions must be taken very cautiously, as the number of studies included in our meta-analysis was very small in comparison with the total number of studies found in the literature that have applied the CY–BOCS.

Finally, it should be noted here that the CY–BOCS was developed based on the OCD definition stated in versions of the *DSM* prior to the *DSM–5* (American Psychiatric Association, 2013). As a consequence, some aspects considered in the fifth edition, such as new symptom categories (e.g., tic-related symptoms) are not yet accounted for by this scale. In addition, the CY–BOCS is an adaptation for children and adolescents from the original Y–BOCS. A new version of the latter is now available (Y–BOCS–II; Goodman, Rasmussen, Price, & Storch, 2006), which includes changes in the scoring and the criteria used to assess the different items. These changes can also be reasonably expected to affect future applications of the CY–BOCS.

Funding

This research was supported by a grant from the Fundación Séneca, Region of Murcia (Spain; Project No. 08650/PHCS/08).

References

References marked by an asterisk indicate studies included in the meta-analysis.

*Abali, O., Nazik, H., Gurkan, K., Unuvar, E., Sidal, M., Ongen, B., … Tuzun, U. (2006). Group A beta hemolytic streptococcal infections and obsessive–compulsive symptoms in a Turkish pediatric population. *Psychiatry and Clinical Neurosciences*, *60*, 103–105.

Abramowitz, J. S., Taylor, S., & McKay, D. (2009). Obsessive compulsive disorder. *Lancet*, *374*, 491–499.

*Adrianzen-Ronceros, C., Pacheco-Armas, Z., Vivar-Cuba, R., & Macciota-Felices, B. (2008). Validez y confiabilidad de la Escala de Yale Brown versión niños y adolescentes (cy-bocs) en el Perú [Validity and reliability of the Yale–Brown Obsessive–Compulsive Scale in Peru]. *Revista Peruana de Pediatría*, *61*, 68–75.

Alaghband-Rad, J., & Hakimshooshtary, M. (2009). A randomized controlled clinical trial of Citalopram versus Fluoxetine in children and adolescents with obsessive–compulsive disorder (OCD). *European Child and Adolescent Psychiatry*, *18*, 131–135.

Alvarenga, P. G., Hounie, A. G., Mercadante, M. T., Diniz, J. B., Salem, M., Spina, G., & Miguel, E. C. (2006). Obsessive–compulsive symptoms in heart disease patients with and without history of rheumatic fever. *Journal of Neuropsychiatry and Clinical Neurosciences*, *18*, 405–408.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

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

Apter, A., Fallon, T. J., King, R. A., Ratzoni, G., Zohar, A. H., Binder, M., … Cohen, D. J. (1996). Obsessive–compulsive characteristics: From symptoms to syndrome. *Journal of the American Academy of Child and Adolescent Psychiatry*, *35*, 907–912.

*Ballesteros-Montero, A. T., & Ulloa-Flores, R. E. (2011). Estudio comparativo de las características clínicas, demográficas y el funcionamiento familiar en niños y adolescentes con trastorno obsessivo-compulsivo leve a moderado vs. grave [Comparative study of the clinical and demographic characteristics and familiar functioning in children and adolescents with mild to moderate vs. severe obsessive–compulsive disorder]. *Salud Mental*, *34*, 121–128.

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

Ben-Dor, D. H., Zimmerman, S., Sever, J., Roz, N., Apter, A., Rehavi, M., & Weizman, A. (2007). Reduced platelet vesicular monoamine transporter density in Tourette's syndrome pediatric male patients. *European Neuropsychopharmacology*, *17*, 523–526.

*Björgvinsson, T., Wetterneck, C. T., Powell, D. M., Chasson, G. S., Webb, S. A., Hart, J., … Stanley, M. A. (2008). Treatment outcome for adolescent obsessive–compulsive disorder in a specialized hospital setting. *Journal of Psychiatric Practice*, *14*, 137–145.

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

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*, 368–385.

Borenstein, M. J., Hedges, L. V., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis.* Chichester, England: Wiley.

*Brynska, A., & Wolanczyk, T. (2005). Epidemiology and phenomenology of obsessive–compulsive disorder in non-referred young adolescents: A Polish perspective. *European Child and Adolescent Psychiatry*, *14*, 319–327.

Chabane, N., Delorme, R., Millet, B., Mouren, M.-C., Leboyer, M., & Pauls, D. (2005). Early-onset obsessive–compulsive disorder: A subgroup with a specific clinical and familial pattern? *Journal of Child Psychology and Psychiatry*, *46*, 881–887.

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

Cubo, E., Fernández-Jaén, A., Moreno, C., Anaya, B., González, M., & Kompoliti, K. (2008). Donepezil use in children and adolescents with tics and

attention-deficit/hyperactivity disorder: An 18-week, single-centre, dose-escalating, prospective, open-label study. *Clinical Therapy*, *30*, 182–189.

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2008). Analysing data and undertaking meta-analyses. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook of systematic reviews of interventions* (pp. 243–296). Chichester, England: Wiley.

Dewrang, P., & Sandberg, A. D. (2011). Repetitive behaviour and obsessive–compulsive features in Asperger syndrome: Parental and self-reports. *Research in Autism Spectrum Disorders*, *5*, 1176–1186.

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.

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement*, *66*, 215–227.

*Flessner, C. A., Allgair, A., Garcia, A., Freeman, J., Sapyta, J., Franklin, M. E., … March, J. (2010). The impact of neuropsychological functioning on treatment outcome in pediatric obsessive–compulsive disorder. *Depression and Anxiety*, *27*, 365–371.

*Franklin, M. E., Sapyta, J., Freeman, J. B., Khanna, M., Compton, S., Almirall, D., … March, J. S. (2011). Cognitive behaviour therapy augmentation of pharmacotherapy in pediatric obsessive–compulsive disorder: The Pediatric OCD Treatment Study II (POTS II) randomized controlled trial. *Journal of the American Medical Association*, *306*, 1224–1232.

*Freeman, J., Flessner, C. A., & García, A. (2011). The Children's Yale–Brown Obsessive Compulsive Scale: Reliability and validity for use among 5 to 8 year olds with obsessive–compulsive disorder. *Journal of Abnormal Child Psychology*, *39*, 877–883.

Gallant, J., Storch, E. A., Merlo, L. J., Ricketts, E. D., Geffken, G. R., Goodman, W. K., & Murphy, T. K. (2008). Convergent and discriminant validity of the Children's Yale–Brown Obsessive Compulsive Scale symptom checklist. *Journal of Anxiety Disorders*, *22*, 1369–1376.

*Geffken, F. R., Storch, E. A., Duke, D. C., Monaco, L., Lewin, A. B., & Goodman, W. K. (2006). Hope and coping in family members of patients with obsessive–compulsive disorder. *Journal of Anxiety Disorders*, *20*, 614–629.

*Geller, D., Petty, C., Vivas, F., Johnson, J., Pauls, D., & Biederman, J. (2007). Examining the relationship between obsessive–compulsive disorder and attention-deficit/hyperactivity disorder in children and adolescents: A familial risk analysis. *Biological Psychiatry*, *61*, 316–321.

*Ginsburg, G. S., Burstein, M., Becker, K. D., & Drake, K. L. (2011). Treatment of obsessive compulsive disorder in young children: An intervention model and case series. *Child and Family Behavior Therapy*, *33*, 97–122.

*Godoy, A., Gavino, A., Valderrama, L., Quintero, C., Cobos, M. P., Casado, Y., … Capafons, J. I. (2011). Estructura factorial y fiabilidad de la adaptación española de la escala obsesivo–compulsiva de Yale–Brown para niños y adolescentes en su versión de autoinforme (CY–BOCS–SR) [Factor structure and reliability of the Spanish adaptation of the Obsessive–Compulsive Scale for children and adolescent self-report version (CY–BOCS–SR)]. *Psicothema*, *23*, 330–335.

Goodman, W. K., Rasmussen, S. A., Price, L. H., & Storch, E. A. (2006). *Y–BOCS–II: Clinical version*. Unpublished manuscript, University of Florida College of Medicine, Gainesville, FL.

*Gorman, D. A., Zhu, H., Anderson, G. M., Davies, M., & Peterson, B. S. (2006). Ferritin levels and their association with regional brain volumes in Tourette's syndrome. *American Journal of Psychiatry*, *163*, 1264–1272.

*Guldeniz-Yucelen, A., Rodopman-Arman, A., Topcuoglu, V., Yanki-Yazgan, M., & Fisek, G. (2006). Interrater reliability and clinical efficacy of Children's Yale–Brown Obsessive–Compulsive Scale in an outpatient setting. *Comprehensive Psychiatry*, *47*, 48–53.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, *41*, 901–916.

Henson, R. K., & 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*, 113–127.

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.

Ivarsson, T., & Valderhaug, R. (2006). Symptom patterns in children and adolescents with obsessive–compulsive disorder (OCD). *Behaviour Research and Therapy*, *44*, 1106–1116.

Ivarsson, T., & Winge-Westholm, C. (2004). Temperamental factors in children and adolescents with obsessive–compulsive disorder (OCD) and in normal controls. *European Child and Adolescent Psychiatry*, *13*, 365–372.

*Keeley, M. L., Geffken, G. R., Ricketts, E., McNamara, J. P. H., & Storch, E. A. (2011). The therapeutic alliance in the cognitive behavioural treatment of pediatric obsessive–compulsive disorder. *Journal of Anxiety Disorders*, *25*, 855–863.

Kim, J.-W., Yoo, H.-J., Soo-Churcl, C., Kang-E, M. H., & Boong-Nyun, K. (2005). Behavioral characteristics of Prader–Willi syndrome in Korea: Comparison with children with mental retardation and normal controls. *Journal of Child Neurology*, *20*, 134–138.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710.

Leach, L. F., Henson, R. K., Odom, L. R., & Cagle, L. S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educational and Psychological Measurement*, *66*, 285–304.

*Lewin, A. B., Bergman, R. L., Peris, T. S., Chang, S., McCracken, J. T., & Piacentini, J. (2010). Correlates of insight among youth with obsessive–compulsive disorder. *Journal of Child Psychology and Psychiatry*, *51*, 603–611.

*Lewin, A. B., Caporino, N., Murphy, T. K., Geffken, G. R., & Storch, E. A. (2010). Understudied clinical dimensions in pediatric obsessive compulsive disorder. *Child Psychiatry and Human Development*, *41*, 675–691.

*Lewin, A. B., Peris, T. S., Bergman, R. L., McCracken, J. T., & Piacentini, J. (2011). The role of treatment in youth receiving exposure-based CBT for obsessive compulsive disorder. *Behaviour Research and Therapy*, *49*, 536–543.

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.

Maina, G., Albert, U., Bogetto, F., & Ravizza, L. (1999). Obsessive–compulsive syndromes in older adolescents. *Acta Psychiatrica Scandinavica*, *100*, 447–450.

*Marrs-Garcia, A., Sapyta, J. J., Moore, P. S., Freeman, J. B., Franklin, M. E., March, J. S., & Foa, E. B. (2010). Predictors and moderators of treatment outcome in the Pediatric Obsessive Compulsive Treatment Study (POTS I). *Journal of American Academy of Child and Adolescent Psychiatry*, *49*, 1024–1033.

Mason, C., Allam, R., & Brannick, M. T. (2007). How to meta-analyze coefficient-of- stability estimates: Some recommendations on Monte Carlo studies. *Educational and Psychological Measurement*, *67*, 765–783.

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.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

*McKay, D., Piacentini, J., Greisberg, S., Graae, F., Jaffer, M., Miller, J., … Yaryura-Tobias, J. A. (2003). The Children's Yale–Brown Obsessive–Compulsive Scale: Item structure in an outpatient setting. *Psychological Assessment*, *15*, 578–581.

Merlo, L. J., Storch, E. A., Murphy, T. K., Goodman, W. K., & Geffken, G. R. (2005). Assessment of pediatric obsessive–compulsive disorder: A critical

review of current methodology. *Child Psychiatry and Human Development*, *36*, 195–214.

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.

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.

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

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 139–162). New York, NY: Russell Sage Foundation.

*The Pediatric OCD Treatment Study (POTS) Team. (2004). Cognitive-behavior therapy, sertraline, and their combination for children and adolescents with obsessive–compulsive disorder. *Journal of the American Medical Association*, *292*, 1969–1976.

*Peris, T. S., Bergman, R. L., Asarnow, J. R., Langley, A., McCracken, J. T., & Piacentini, J. (2010). Clinical and cognitive correlates of depressive symptoms among youth with obsessive compulsive disorder. *Journal of Clinical Child and Adolescent Psychology*, *39*, 616–626.

*Peris, T. S., Bergman, R. L., Langley, A., Chang, S., McCracken, J. T., & Piacentini, J. (2008). Correlates of accommodation of pediatric obsessive–compulsive disorder: Parent, child, and family characteristics. *Journal of American Academy of Child and Adolescent Psychiatry*, *47*, 1173–1181.

*Peterson, B. S., Prakash, T., Kane, M. J., Scahill, L., Zhang, H., Bronen, R., … Staib, L. (2003). Basal ganglia volumes in patients with Gilles de la Tourette syndrome. *Archives of General Psychiatry*, *60*, 415–424.

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.

Rapoport, J. L., Inoff-Germain, G., Weissman, M. M., Greenwald, S., Narrow, W. E., Jensen, P. S., … Canino, G. (2000). Childhood obsessive–compulsive disorder in the NIMH MECA study: Parent versus child identification of cases. Methods for the epidemiology of child and adolescent mental disorders. *Journal of Anxiety Disorders*, *14*, 535–548.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–315). New York, NY: Russell Sage Foundation.

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

Rouse, S. V. (2007). Using reliability generalization methods to explore measurement error: An illustration using the MMPI–2 PSY–5 scales. *Journal of Personality Assessment*, *88*, 264–275.

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.

*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 American Academy of Child and 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.

Steinberg, K., & Schuch, B. (2002). Classification of obsessive–compulsive disorder in childhood and adolescence. *Acta Psychiatrica Scandinavica*, *106*, 97–102.

*Steinberg, T., Baruch, S. S., Harush, A., Dar, R., Woods, D., Piacentini, J., & Apter, A. (2010). Tic disorders and the premonitory urge. *Journal of Neural Transmission*, *117*, 277–284.

Storch, E. A., Benito, K., & Goodman, W. (2011). Assessment scales for obsessive–compulsive disorder. *Neuropsychiatry*, *1*, 243–250.

*Storch, E. A., Caporino, N. E., Morgan, J. R., Lewin, A. B., Rojas, A., Brauer, L., … Murphy, T. K. (2011). Preliminary investigation of web-camera delivered cognitive-behavioral therapy for youth with obsessive–compulsive disorder. *Psychiatry Research*, *189*, 407–412.

*Storch, E. A., Geffken, G. F., Merlo, L. J., Jacob, M. L., Murphy, T. K., Goodman, W. K., … Grabill, K. (2007). Family accommodation in pediatric obsessive–compulsive disorder. *Journal of Clinical Child and Adolescent Psychology*, *36*, 207–216.

*Storch, E. A., Geffken, G. R., Merlo, L. J., Mann, G., Duke, D., Munson, M., … Goodman, W. K. (2007). Family-based cognitive-behavioral therapy for pediatric obsessive–compulsive disorder: Comparison of intensive and weekly approaches. *Journal of American Academy of Child and Adolescent Psychiatry*, *46*, 469–478.

*Storch, E. A., Larson, M. J., Muroff, J., Caporino, N., Geller, D., Reid, J. M., … Murphy, T. K. (2010). Predictors of functional impairment in pediatric obsessive–compulsive disorder. *Journal of Anxiety Disorders*, *24*, 275–283.

*Storch, E. A., Ledley, D. R., Lewin, A. B., Murphy, T. K., Johns, N. B., Goodman, W. K., & Geffken, G. R. (2006). Peer victimization to children with obsessive–compulsive disorder: Relations with symptoms of psychopathology. *Journal of Clinical Child and Adolescent Psychology*, *35*, 446–455.

*Storch, E. A., Lehmkuhl, H., Pence, S. L., Jr., Geffken, G. R., Ricketss, E., Storch, J. F., & Murphy, T. K. (2009). Parental experiences of having a child with obsessive–compulsive disorder: Associations with clinical characteristics and caregiver adjustment. *Journal of Child and Family Studies*, *18*, 249–258.

*Storch, E. A., Lewin, A. B., De Nadai, A. S., & Murphy, T. K. (2010). Defining treatment response and remission in obsessive–compulsive disorder: A signal detection analysis of the Children's Yale–Brown Obsessive Compulsive Scale. *Journal of American Academy Child and Adolescent Psychiatry*, *49*, 708–717.

*Storch, E. A., Merlo, L. J., Keeley, M. L., Grabill, K., Milsom, V. A., Geffken, G. R., … Goodman, W. K. (2008). Somatic symptoms in children and adolescents with obsessive–compulsive disorder: Associations with clinical characteristics and cognitive-behavioral therapy response. *Behavioural and Cognitive Psychotherapies*, *36*, 283–297.

*Storch, E. A., Merlo, L. J., Larson, M. J., Geffken, G. R., Lehmkuhl, H. D., Jacob, M. … Goodman, W. K. (2008). Impact of comorbidity on cognitive-behavioral therapy response in pediatric obsessive–compulsive disorder. *Journal of American Academy of Child and Adolescent Psychiatry*, *47*, 583–592.

*Storch, E. A., Muroff, J., Lewin, A. B., Geller, D., Ross, A., McCarthy, K., … Steketee, G. (2011). Development and preliminary psychometric evaluation of the Children's Saving Inventory. *Child Psychiatry and Human Development*, *42*, 166–182.

*Storch, E. A., Murphy, T. K., Adkins, J. W., Lewin, A. B., Geffken, G. R., Johns, N. B., … Goodman, W. K. (2006). The Children's Yale–Brown Obsessive–Compulsive Scale: Psychometric properties of child- and parent-report formats. *Journal of Anxiety Disorders*, *20*, 1055–1070.

*Storch, E. A., Murphy, T. K., Bagner, D. M., Johns, N. B., Baumeister, A. L., Goodman, W. K., & Geffken, G. R. (2006). Reliability and validity of the Child Behavior Checklist Obsessive–Compulsive Scale. *Journal of Anxiety Disorders*, *20*, 473–485.

*Storch, E. A., Murphy, T. K., Geffken, G. R., Bagner, D. M., Soto, O., Sajid, M., … Goodman, W. K. (2005). Factor analytic study of the Children's Yale–Brown Obsessive–Compulsive Scale. *Journal of Clinical Child and Adolescent Psychology*, *34*, 312–319.

*Storch, E. A., Murphy, T. K., Geffken, G. R., Mann, G., Adkins, J., Merlo, L. J., … Goodman, W. K. (2006). Cognitive-behavioral therapy for PANDAS-related obsessive–compulsive disorder: Findings from a preliminary waitlist controlled open trial. *Journal of American Academy of Child and Adolescent Psychiatry*, *45*, 1171–1178.

*Storch, E. A., Murphy, T. K., Geffken, G. R., Sajid, M., Allen, P., Roberti, J. W., & Goodman, W. K. (2005). Reliability and validity of the Yale Global Tic Severity Scale. *Psychological Assessment*, *17*, 486–491.

*Storch, E. A., Murphy, T. K., Geffken, G. R., Soto, O., Sajid, M., Allen, P., . . . Goodman, W. K. (2004a). Further psychometric properties of the Tourette's Disorder Scale–parent rated version (TODS–PR). *Child Psychiatry and Human Development*, *35*, 107–120.

*Storch, E. A., Murphy, T. K., Geffken, G. R., Soto, O., Sajid, M., Allen, P., . . . Goodman, W. K. (2004b). Psychometric evaluation of the Children's Yale–Brown Obsessive–Compulsive Scale. *Psychiatry Research*, *129*, 91–98.

*Storch, E. A., Stigge-Kaufman, D., Marien, W. E., Sajid, M., Jacob, M. L., Geffken, G. R., . . . Murphy, T. K. (2008). Obsessive–compulsive disorder in youth with and without a chronic tic disorder. *Depression and Anxiety*, *25*, 761–767.

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.

Taylor, S. (2011). Early versus late onset obsessive–compulsive disorder: Evidence for distinct subtypes. *Clinical Psychology Review*, *31*, 1083–1100.

Termine, C., Balottin, U., Rossi, G., Maisano, F., Salini, S., Di Nardo, R., & Lanzi, G. (2006). Psychopathology in children and adolescents with Tourette's syndrome: A controlled study. *Brain Development*, *28*, 69–75.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174–195.

Thomsen, P. H. (1994). Obsessive–compulsive disorder in children and adolescents: A study of phenomenology and family functioning in 20 consecutive Danish cases. *European Child and Adolescent Psychiatry*, *3*, 29–36.

*Ulloa, R. E., de la Peña, F., Higuera, F., Palacios, H., Nicolini, H., & Ávila, J. M. (2004). Validity and reliability of the Spanish version of Yale–Brown Obsessive–Compulsive Rating Scale for children and adolescents. *Actas Españolas de Psiquiatría*, *32*, 216–221.

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.

Vassar, M., & Bradley, G. (2010). A reliability generalization study of coefficient alpha for the Life Orientation Test. *Journal of Personality Assessment*, *92*, 362–370.

Vassar, M., & Crosby, J. W. (2008). A reliability generalization study of coefficient alpha for the UCLA Loneliness Scale. *Journal of Personality Assessment*, *90*, 601–607.

Verhaak, L. M., & De Haan, E. (2007). Cognitions in children with OCD: A pilot study for age specific relations with severity. *European Child and Adolescent Psychiatry*, *16*, 353–361.

Wang, H.-S., & Kuo, M.-F. (2003). Sonographic lenticulostriate vasculopathy in infancy with tic and other neuropsychiatric disorders developed after 7 to 9 years of follow-up. *Brain Development*, *25*(Suppl. 1), S43–S47.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

*Ye, H. J., Rice, K. G., & Storch, E. A. (2008). Perfectionism and peer relations among children with obsessive–compulsive disorder. *Child Psychiatry and Human Development*, *39*, 415–426.

Zohar, A. H. (1999). The epidemiology of obsessive–compulsive disorder in children and adolescents. *Child and Adolescent Psychiatric Clinics of North America*, *8*, 445–460.

APPENDIX.—Studies included in the meta-analysis.

Study	N_{TOTAL}	$\hat{\alpha}_i$	ICC_i	$\hat{\kappa}_i$	N_{AGREE}	Test Version	Score *M*	Score *SD*	Age *M*	Age *SD*	Clinical Sample	Disorder
Abali et al. (2006)	31	—	.96	—	20	Other	9.70	7.60	8.00	2.90	Yes	Other
Adrianzen-Ronceros et al. (2008)	46	.87	—	—	—	Other	—	—	13.30	2.50	Yes	OCD
Ballesteros-Montero & Ulloa-Flores (2011)	60	.87	—	—	—	Other	22.76	8.66	12.57	2.91	No	—
Björgvinsson et al. (2008)	20	.87	.84	—	20	Original	23.90	8.60	15.30	—	Yes	OCD
Brynska & Wolanczyk (2005)	148	.91	—	—	—	Other	—	—	14.40	—	No	OCD
Flessner et al. (2010)	63	—	.88	—	63	Original	—	—	11.80	2.60	Yes	OCD
Franklin et al. (2011)	124	—	.97	—	12	Original	26.29	5.05	13.60	2.77	Yes	OCD
Freeman et al. (2011)	42	.72	.79	—	42	Original	22.40	4.30	6.70	1.20	Yes	OCD
Geffken et al. (2006)	17	.85	—	—	—	Original	—	—	—	—	Yes	OCD
Geller et al. (2007)	500	—	—	.87	500	Original	—	—	—	—	Yes	Other
Ginsburg et al. (2011)	7	.78	—	—	—	Original	30.57	5.26	6.00	1.93	Yes	OCD
Godoy et al. (2011)	1,706	.58	.66	—	263	Other	8.46	5.40	13.46	1.62	No	—
Gorman et al. (2006)	42	—	.90	—	42	Original	—	—	—	—	—	—
Guldeniz-Yucelen et al. (2006)	19	.77	.89	—	19	Other	21.05	7.77	14.00	2.25	Yes	OCD
Keeley et al. (2011)	25	.81	—	—	—	Other	25.73	4.59	13.16	2.69	Yes	OCD
Lewin, Bergman, et al. (2010)	71	.74	—	—	—	Original	24.90	4.74	11.76	2.45	Yes	OCD
Lewin, Caporino, et al. (2010)	89	.81	—	—	—	Original	24.16	5.07	12.60	2.80	Yes	OCD
Lewin et al. (2011)	49	.77	—	—	—	Original	—	—	12.00	2.60	Yes	OCD
Marrs-García et al. (2010)	100	—	.81	—	62	Original	—	—	11.70	2.70	Yes	OCD
McKay et al. (2003)	233	.95	—	—	—	Original	—	—	10.80	3.19	Yes	OCD
Peris et al. (2008)	65	.73	—	—	—	Original	25.03	4.73	12.25	—	Yes	OCD
Peris et al. (2010)	71	.73	.93	—	71	Original	24.87	4.67	12.17	2.48	Yes	Other
Peterson et al. (2003)	173	—	.90	—	173	Original	—	—	—	—	Yes	Other
The POTS Team (2004)	112	—	.81	—	112	Original	24.38	4.22	11.78	2.78	Yes	OCD
Scahill et al. (1997)	65	.87	.84	.42	24	Original	19.80	7.55	12.10	2.66	Yes	OCD
Steinberg et al. (2010)	40	.89	.85	—	20	Other	—	—	11.05	2.05	Yes	Other
Storch et al. (2004b)	61	.90	.79	—	37	Original	21.87	7.69	10.33	3.13	Yes	OCD
Storch et al. (2004a)	67	.84	—	—	—	Original	19.00	12.20	9.75	2.25	Yes	Other
Storch, Murphy, Geffken, Sajid, et al. (2005)	28	.94	—	—	—	Original	—	—	10.47	2.51	Yes	Other
Storch, Murphy, Geffken, Bagner, et al. (2005)	82	.76	—	—	—	Original	—	—	10.40	3.00	Yes	OCD
Storch, Murphy, Adkins, et al. (2006)	53	.89	—	—	—	Original	19.90	8.60	11.30	2.40	Yes	OCD
Storch, Ledley, et al. (2006)	52	.90	—	—	—	Original	—	—	11.30	2.30	Yes	OCD
Storch, Murphy, Geffken, et al. (2006)	7	—	—	.97	6	Original	28.00	4.60	11.10	1.40	Yes	Other
Storch, Murphy, Bagner, et al. (2006)	42	.87	—	—	—	Original	—	—	10.50	3.30	Yes	Other
Storch, Stigge-Kaufman, et al. (2008)	74	.90	.79	—	74	Original	26.55	6.68	9.65	2.31	Yes	OCD
Storch, Geffken, Merlo, Jacob, et al. (2007)	57	.89	—	.97	20	Original	26.13	7.26	12.99	2.54	Yes	OCD
Storch, Geffken, Merlo, Mann, et al. (2007)	40	—	—	.96	20	Original	25.65	5.74	13.30	2.70	Yes	OCD
Storch, Merlo, Keeley, et al. (2008)	85	—	—	.96	85	Original	—	—	12.85	3.02	Yes	OCD
Storch, Merlo, Larson, et al. (2008)	96	—	—	.96	20	Original	28.75	5.01	13.40	3.43	Yes	OCD
Storch, Lewin, et al. (2010)	109	—	—	.97	36	Original	—	—	—	—	Yes	OCD
Storch, Caporino, et al. (2011)	31	.71	—	—	—	Original	23.39	3.82	11.09	2.66	Yes	OCD
Storch, Larson, et al. (2010)	99	.86	—	—	—	Original	—	—	12.84	2.81	Yes	OCD
Storch et al. (2009)	62	.86	—	.98	18	Original	—	—	12.56	3.57	Yes	OCD
Storch, Muroff, et al. (2011)	123	.88	.99	—	31	Original	23.00	5.90	13.00	2.90	Yes	OCD
Ulloa et al. (2004)	28	.87	.94	—	28	Other	16.50	9.80	12.10	2.70	Yes	OCD
Ye et al. (2008)	31	.91	—	—	—	Original	19.16	9.43	11.77	2.59	Yes	OCD

Note. N_{TOTAL} = total sample size ; $\hat{\alpha}_i$ = coefficient alpha; ICC_i = intraclass correlation estimating the interrater agreement reliability; $\hat{K}_i$ = kappa coefficient estimating the interrater agreement reliability; N_{AGREE} = sample size employed to estimate the interrater agreement reliability; score *M* = mean of the CY–BOCS total score; score *SD* = standard deviation of the CY–BOCS total score; age *M* = mean age of the sample (in years); age *SD* = standard deviation of the age of the sample (in years); OCD = obsessive–compulsive disorder. Some studies also reported some other kinds of reliability that have not been included in this table. One study (Dewrang & Sandberg, 2011) reported neither a coefficient alpha, nor an intraclass correlation, nor a kappa coefficient, but a Spearman correlation, so that it was not included in this table.