Psicothema

Psicothema 2021, Vol. 33, No. 3, 481-489
doi: 10.7334/psicothema2020.455

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

A Reliability Generalization Meta-analysis of the Dimensional Obsessive-Compulsive Scale

Rubén López-Nicolás[1], María Rubio-Aparicio[2], Carmen López-Ibáñez[1], and Julio Sánchez-Meca[1]
[1] Universidad de Murcia, and [2] Universidad de Alicante

Abstract

Background: The Dimensional Obsessive-Compulsive Scale (DOCS) is a well-established tool for assessing obsessive-compulsive symptomatology. A reliability generalization meta-analysis was conducted to estimate the average reliability of DOCS scores and how reliability estimates vary according to the composition and variability of samples, to identify study characteristics that can explain its variability, and to estimate the reliability induction rate. **Method:** A literature search produced 86 studies that met the inclusion criteria. **Results:** For the DOCS total scores, an average alpha coefficient of .925 was found (95% CI [.920,.931]), as well as mean alphas of .881, .905, .913, and .914 for Contamination, Responsibility, Unacceptable Thoughts, and Symmetry subscales, respectively. Moderator analysis showed that internal consistency fell significantly the more clinical and subclinical participants there were in the sample, and the larger the mean score in the sample for the total scores. The most important moderator variables for the subscales were the standard deviation and the mean of the scores. **Conclusions:** The DOCS scores exhibited excellent internal consistency reliability for both total score and subscale scores and DOCS is suitable both for research and clinical purposes.

Keywords: Meta-analysis; reliability generalization; obsessive-compulsive disorder; Dimensional Obsessive-Compulsive Scale; Cronbach's alpha coefficient; test-retest.

Resumen

Meta-análisis de Generalización de la Fiabilidad de la Dimensional Obssesive-Compulsive Scale. **Antecedentes:** la Dimensional Obsessive-Compulsive Scale (DOCS) (Abramowitz, 2010) es un instrumento para la evaluación del TOC. En este estudio se llevó a cabo un estudio de generalización de la fiabilidad de la DOCS para estimar un coeficiente alfa medio y analizar la heterogeneidad de estos y el influjo de distintas variables moderadores. **Método:** se realizó una búsqueda bibliográfica de 86 estudios incluibles. **Resultados:** para la escala total, se estimó un coeficiente alfa medio de .925, así como para sus subescalas: Contaminación .881, Responsabilidad con respecto al daño .905, Pensamientos inaceptables .913 y Simetría .913. Los análisis de moderadores revelaron que la consistencia interna disminuyó significativamente a mayor porcentaje de participantes clínicos o subclínicos en la muestra, así como a mayor puntuación media de la muestra para las puntuaciones totales; para las subescalas la desviación típica y la media de las puntuaciones fueron los moderadores más relevantes. **Conclusiones:** las puntuaciones de la DOCS muestran una excelente fiabilidad por consistencia interna, tanto para la escala global como para las subescalas, pudiendo usarse tanto para fines clínicos como de investigación.

Palabras clave: meta-análisis; generalización de la fiabilidad; trastorno obsesivo-compulsivo; Dimensional Obsessive-Compulsive Scale; alfa de Cronbach; test-retest.

Obsessive-compulsive disorder (OCD) is a disease characterized by obsessive thoughts, compulsive behaviors or both. Obsessions are recurrent as well as ego dystonic thoughts that cause distress. Compulsions refer to repetitive behaviors that a person with OCD feels the urge to carry out in response to obsessive thoughts (American Psychiatric Association [APA], 2013). Lifetime and 12-month prevalence of OCD in adults are estimated at about 2.3% and 1.2%, respectively (Ruscio et al., 2010).

In recent years, OCD has received renewed interest due to the inclusion of a new diagnostic category in the DSM-V. While in the DSM-IV-TR OCD was included in the 'Anxiety disorders 'category, the latest edition of this diagnostic manual has added the 'Obsessive-compulsive and related disorders 'category, which includes OCD, 'body dysmorphic disorder', 'hoarding disorder', 'trichotillomania' and 'excoriation disorder'.

OCD patients can manifest very different types of obsessions (e.g. sexual, religious, contamination fears) as well as various compulsive behaviors to cope with obsessions, therefore OCD is considered a multidimensional disorder (Mataix-Cols et al., 2005). The heterogeneity exhibited in OCD clinical manifestations is an important point to be considered when assessing this disorder.

The *Dimensional Obsessive-Compulsive Scale* (DOCS) is a measurement tool developed by Abramowitz et al. (2010) that intends to solve some of the limitations of other scales to assess obsessive-compulsive symptomatology, that typically have a larger number of items, with checklists assessing specific symptoms, being time inefficient and misinterpreting the range of symptoms with severity.

The DOCS is a 20-item scale structured in four dimensions or subscales with 5 items each: *Contamination*, *Responsibility for*

Received: December 7, 2020 • Accepted: February 19, 2021
Corresponding author: Julio Sánchez-Meca
Facultad de Psicología
Universidad de Murcia
30100 Murcia (Spain)
email: jsmeca@um.es

Harm, *Unacceptable Thoughts*, and *Symmetry*. Distress generated in the patient is assessed on five parameters for each dimension: time occupied by obsessions and compulsions, avoidance behavior, associated distress, functional interference, and difficulty disregarding obsessions and refraining from compulsions (Abramowitz et al., 2010). Each parameter is scored on a Likert-type scale with five categories (0: null distress; 4: high distress), thus each subscale provides scores ranging from 0-20 and a DOCS total score can also be calculated ranging from 0-80. In its original validation study, Abramowitz et al. (2010) found excellent internal consistency of the DOCS total score, satisfactory convergent and discriminant validity, and the factor structure was in accordance with the theoretical proposal initially established. A summary of the psychometric properties evidenced in the validation studies of original DOCS and its validation version is available at: https://osf.io/cnpfa/

As classical test theory states, reliability is not an inherent property of the test, but of the test scores obtained in a given application. Reliability of test scores changes according to the composition and variability of the sample and of the context of application (Crocker & Algina, 1986; Streiner & Norman, 2008). This is why many scientific organizations and journals recommend empirical studies report reliability estimates with the data at hand and avoid the malpractice of inducing reliability from previous applications of the test (Appelbaum et al., 2018; Heldref Foundation [HF],1997; Thompson, 1994; Vacha-Haase et al., 2000; Wilkinson & the APA Task Force for Statistical Inference [APA TFSI], 1999). Consequently, it is important to carry out studies to examine in what extent reliability of test scores varies as it is applied to different samples of participants and other potential moderators. A reliability generalization (RG) meta-analysis is a special kind of meta-analysis aiming to investigate how measurement error of test scores changes from one application to the next and to identify which study and sample characteristics are statistically associated to variability of reliability estimates (Henson & Thompson, 2002; Rodríguez & Maeda, 2006; Sánchez-Meca et al., 2013; Vacha-Haase, 1998).

To date, several RG meta-analyses have been conducted on different scales that assess obsessive-compulsive symptomatology in adults, such as the *Maudsley Obsessive-Compulsive Inventory* (MOCI, Sánchez-Meca et al., 2011), the *Yale-Brown Obsessive-Compulsive Scale* (Y-BOCS, López-Pina et al., 2015), the *Padua Inventory of Obsessions and Compulsions* (PI, Sánchez-Meca et al., 2017), the *Padua Inventory-Washington State University Revision* (PI-WSUR, Rubio-Aparicio et al., 2020), and the *Padua Inventory-Revised of Obsessive-Compulsive Symptoms* (PI-R, Núñez-Núñez et al., 2020). To the best of our knowledge, an RG meta-analysis on the DOCS scores has not yet been conducted.

An RG meta-analysis was conducted on the DOCS with the following objectives: (a) estimate the average reliability of the DOCS scores, both total score and subscales; (b) assess whether reliability estimates of the DOCS were heterogeneous; (c) identify both study characteristics as well as those of samples that can be statistically associated to reliability estimates; (d) propose an explanatory/predictive model of reliability estimates capable of explaining a large proportion of reliability estimate variance, and (e) estimate the reliability induction rate of the DOCS. In addition, we aimed to compare the reliability of the DOCS scores with those of other scales that assess obsessive-compulsive symptomatology. Internal consistency and temporal stability were the kinds of reliability investigated in this meta-analysis.

Method

Review methods and reporting were performed according to the *Reliability Generalization Meta-Analysis* (REGEMA) guidelines (Sánchez-Meca et al., 2021, May). REGEMA checklist for the present meta-analysis is available at https://osf.io/hevuc/

Participants

Study Selection Criteria

Studies were required to fulfil the following inclusion criteria: (a) empirical studies applying the DOCS or an adaptation of this scale maintaining the original structure; (b) studies that reported any reliability estimate with data from the study sample; (c) samples of any target population were accepted (community, clinical or subclinical); (d) paper had to be written in English or Spanish, and (e) published studies were accepted in this meta-analysis.

Study Search Strategy

Relevant articles were identified by systematically searching the following databases: PubMed, PsycInfo and ProQuest, from data inception until June 2020. The search strategy included the keyword "Dimensional Obsessive-Compulsive Scale" to be found anywhere in the full text. No search limits were set. Google Scholar was consulted to identify studies (published or not) that could fulfil the selection criteria.

In the first screening stage, we reviewed titles and abstracts from the search hits, we then full-text reviewed the remaining articles. Figure 1 displays the flowchart illustrating the selection and inclusion process. We initially identified 660 references from which 191 were duplicated and 210 were excluded for various reasons. We full-text reviewed 259 references, from which 103 did not apply the DOCS. Of the remaining 156 references that did apply the DOCS, 119 (76.28%) reported any reliability estimate with the data at hand, whereas 37 (23.71%) induced reliability from previous applications of the test or omitted any reference to the reliability of the DOCS scores.

Instruments

A coding form was developed to extract relevant study characteristics and reliability estimates. The coding form is available at https://osf.io/jghpk/ and contains information about all coded variables.

Procedure

Two authors doubly coded a random sample of 20 studies independently aiming to assess the reliability of the data extraction process. We computed kappa (κ) coefficients for categorical variables and intraclass correlation coefficients (ICC) for continuous variables. Results showed a highly satisfactory agreement overall: κs varying between .86 and 1 (mean = .980) and ICCs varying between .90 and 1 (mean = .992). Discrepancies between coders were resolved by consensus.

Data Analysis

Separate meta-analyses were carried out for the global scale and for each of the four subscales. In addition, separate meta-

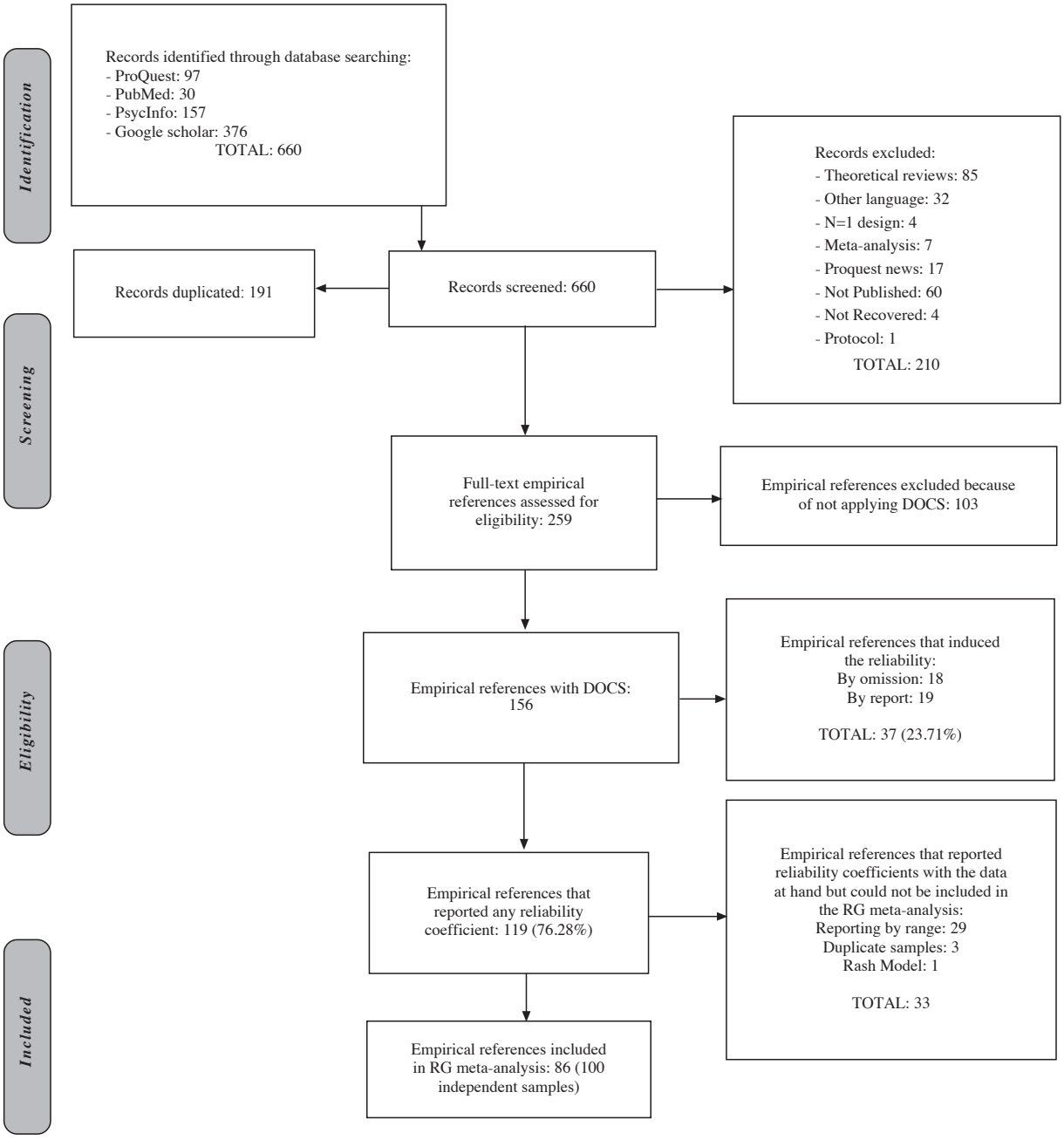

Figure 1. Flow diagram of the searching, screening, and selection process of the studies included in this RG meta-analysis

analyses were conducted for alphas and retest coefficients. A transformation of the alpha coefficients proposed by Bonett (2002) was applied before its statistical integration, in order to normalize its distribution and stabilize variances: $L_i = Ln(1 - |\alpha_i|)$, α_i being the coefficient alpha for each study, *Ln* being the natural logarithm, and L_i being the transformed coefficient. Fisher's Z transformation was applied for retest coefficients. In addition, as a sensitivity analysis, we conducted all meta-analyses both with untransformed and transformed alpha coefficients (Sánchez-Meca et al., 2013).

Random-effects models were assumed to compute summary statistics of reliability coefficients, weighting the coefficients by the inverse variance. Between-studies variance was estimated by the restricted maximum likelihood (REML) method. The 95% confidence limits were computed according to the improved method proposed by Hartung and Knapp (2001; see also Sánchez-Meca & Marín-Martínez, 2008).

The heterogeneity exhibited by the reliability coefficients was assessed computing the *Q* statistic and the I^2 index (Cooper et al., 2019). For the I^2 index, values around 25%, 50%, and 75% were considered as reflecting low, moderate, and large heterogeneity, respectively (Higgins & Thompson, 2002; Huedo-Medina et al., 2006). If significant heterogeneity was found among the reliability coefficients ($I^2 > 25\%$), then moderator analyses were carried out to identify study characteristics that explained heterogeneity.

These analyses were accomplished by applying mixed-effects meta-regression models for the continuous variables and subgroup analyses for the categorical variables with the improved method proposed by Knapp and Hartung (2003; see also López-López et al., 2013; Viechtbauer et al., 2015). All statistical analyses were conducted using the metafor package in R (Viechtbauer, 2010) and the script analysis is available at https://osf.io/djb4h/. Two types of reliability induction were considered (Shields & Caruso, 2004): reliability induction 'by omission' (i.e., to omit any reference to the reliability of the test scores) and reliability induction 'by report' (i.e., to induce reliability of test scores from a previous application of the test).

Results

References from 119 studies that reported reliability coefficients with the data at hand are available at https://osf.io/w56mz/. Of these 119, 33 could not be included in the RG meta-analysis for the following reasons: (a) reporting a range of reliability coefficients instead of genuine values (29 studies); (b) using a duplicate sample of participants (3 studies); (c) and finally, one of these reported reliability under the Rash model. The remaining 86 studies were included in our RG meta-analysis. As several studies reported reliability coefficients for two or more different samples, our study database included a total of 100 independent samples.

The total sample size was N = 27,932 participants, with range 16-2,636 (mean = 279; SD = 346.7). Most studies were carried out in North America (67%), followed by Europe (19%), Oceania (7%), Asia (6%) and Central America (1%). The database used in this study is openly available at https://osf.io/qk4hz/

Internal Consistency Reliability

Table 1 presents the main results for the total scale and for each of the four subscales of the DOCS. As results obtained with the transformed and untransformed coefficients were quite similar, only those from the untransformed coefficients were presented. Figure 2 displays a forest plot of the alpha coefficients of the total scores.

The forest plots of the alpha coefficients for the subscale scores of each study included in the meta-analysis are available at https://osf.io/jhm6u/

The 72 studies that reported an alpha coefficient of the total test score estimated a mean alpha coefficient of 0.925 (95% CI [0.920, 0.931]), ranging from 0.800 to 0.970. For the *Contamination* subscale, 58 studies were included estimating a mean alpha coefficient of 0.881 (95% CI [0.863, 0.899]) with a range from 0.610 to 0.970. For the *Responsibility* subscale, 50 studies were included, estimating a mean alpha coefficient of 0.905 (95% CI [0.893, 0.917]), with a range from 0.790 to 0.960. For *Unacceptable Thoughts* subscale, 51 studies were included, estimating a mean alpha coefficient of 0.913 (95% CI [0.904, 0.922]), with a range from 0.830 to 0.960. Finally, for the *Symmetry* subscale, 49 studies were included estimating a mean alpha coefficient of 0.914 (95% CI [0.906, 0.922]), with a range from 0.850 to 0.960.

Table 1
Mean alpha coefficients, 95% confidence intervals, and heterogeneity statistics for the DOCS total score and the four subscales

Total scale/subscale			95% CI			
	k	α_+	LL	UL	Q	I^2
Total scale	72	.925	.920	.931	567.646**	91.34
Contamination	58	.881	.863	.899	2475.880**	98.32
Responsibility	50	.905	.893	.917	996.208**	96.00
Unacceptable thoughts	51	.913	.904	.922	641.066**	93.97
Symmetry	49	.914	.906	.922	481.511**	91.83

Note: K = number of studies; α_+ = mean coefficient alpha; LL and UL= lower and upper limits of the 95% confidence interval for α_+; Q = Cochran's heterogeneity Q statistic; I^2 = heterogeneity index.**p<.0001

As shown in Table 2, large heterogeneity was found among the reliability coefficients for the total test score and subscales, with I^2 indices larger than 90% in all cases and Q statistics reaching statistical significance.

Temporal Stability Reliability

Only 4 studies reported retest coefficients of the total score and the subscales (N = 580). For total score the average retest coefficient was 0.788 (95% CI [0.124, 0.965]), for *Contamination* 0.633 (95% CI [0, 0.918]), for *Responsibility* 0.703 (95% CI [0.304, 0.892]), for *Unacceptable Thoughts* 0.627 (95% CI [0.141, 0.869]), and for *Symmetry* 0.688 (95% CI [0.115, 0.918]). Heterogeneity was large in all cases (I^2 > 90%).

Analysis of Moderator Variables

Due to the large heterogeneity found among the reliability estimates, analyses of moderator variables were performed. Aiming to guarantee normality of reliability estimate distribution, the analyses of moderators were performed applying the Bonett transformation (2002). To facilitate interpretation of results, the mean alpha coefficients, their confidence limits, and slope estimates obtained with Bonett's transformation were back-transformed to the original metric of alpha coefficient.

DOCS Total Score

Table 2 shows the results of the simple meta-regression analyses applied on the transformed alpha coefficients for the DOCS total score for each continuous moderator variable. From the moderators analysed, the total mean score and the percentage of patients diagnosed with OCD yielded a negative statistically significant relationship with alpha coefficients (p = .002 and p = .003, respectively), with a 13% and 16% of variance explained for both predictors, respectively.

Table 3 presents the results of the weighted ANOVAs applied on the transformed alpha coefficients for the DOCS total score for each categorical moderator variable. Statistically significant differences were found when comparing the mean alpha coefficients grouped by the target population (p < .001), with 36% of variance accounted for. Studies with a mixed target population in particular, exhibited the highest mean alpha coefficient (α_+ = .947). This is to be expected as this mixture of participants increases variability exhibited by test scores. For studies whose target population were community or undergraduate, the mean alpha coefficients were α_+ = .935 and α_+ = .921, respectively, with studies whose target population were clinical or subclinical being those exhibiting lowest mean alpha coefficients (α_+ = .911).

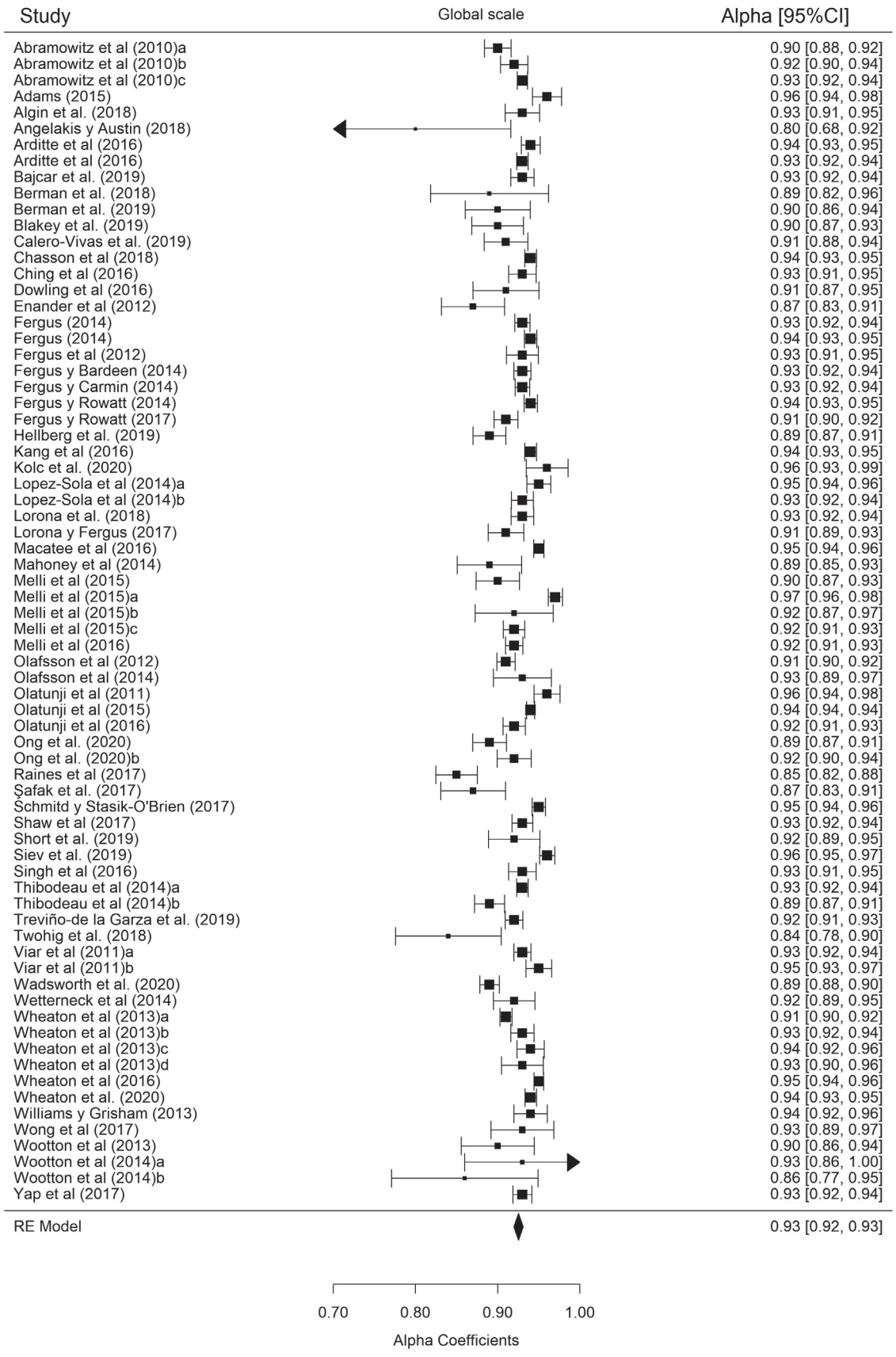

Figure 2. Forest plot of the alpha coefficients reported in the studies on the DOCS total score

Table 2
Results of the simple meta-regressions applied on alpha coefficients for DOCS total score, taking continuous moderator variables as predictors

Predictor variable	k	b_i	F	p	Q_E	R^2
Mean total score	60	-0.0125	6.164	.002	419.30**	.13
SD of total score	59	0.0218	2.118	.151	438.22**	.01
Sample size	72	0.0001	0.995	.322	632.82**	0.0
Mean age (years)	69	0.0072	1.568	.215	529.20**	.03
SD of age (years)	69	0.0074	0.738	.393	545.23**	0.0
Gender (% male)	70	0.0035	0.953	.333	571.15**	0.0
Ethnicity (% Caucasian)	40	-0.0009	0.179	.675	395.80**	0.0
Year of the study	72	-0.0166	1.383	.244	638.94**	0.0
Disorder (% OCD)	65	-0.0030	9.809	.003	512.03**	.16

Note: k = number of studies; b_j= unstandardized regression coefficient; F = Knapp-Hartung's statistic for testing the significance of the predictor (the degrees of freedom for this statistic are 1 for the numerator and $k - 2$ for the denominator); p = probability level for the F statistic; Q_E = statistic for testing the model misspecification; R^2 = proportion of variance accounted for by the predictor. **p<.0001

Although several moderators reached a statistically significant relationship with the alpha coefficients for the total scale, statistics for testing the model misspecification (Q_E for meta-regressions and Q_W for ANOVAs) suggested the presence of residual variability among alpha coefficients. Thus, the next step consisted of fitting a multiple meta-regression model to identify characteristics of studies most relevant to explaining variability among alpha coefficients for the total scale. The predictors included in the model were selected according to results found in the simple meta-regressions previously performed. In particular, predictors that yielded a statistically significant result and a percentage of variance explained above 10% ($R^2 > .10$) were included in the multiple meta-regression model. Thus, for the DOCS total score, the explanatory model included two predictors: total mean score and percentage of patients diagnosed with OCD. Due to missing values in some variables, the number of studies included in the model was $k = 54$. Results are shown in Table 4.

The full model did not reach a statistically significant relationship with the alpha coefficients ($p = .079$), and none of the predictors included in the model showed a statistically significant relationship these either. The full model explained a 10% of observed variance in alpha coefficients. This can be explained by the underlying relationship between the two variables, as samples with a higher total mean score had a higher number of patients diagnosed with OCD ($r_{xy} = .792, p < .001$).

DOCS Subscales

The moderator analyses applied on the transformed alpha coefficients for the *Contamination*, *Responsibility*, *Unacceptable thoughts*, and *Symmetry* subscale scores showed that the most relevant moderator variables were the standard deviation and the mean of each subscale scores. In all cases, the larger the standard deviation of the subscale scores, the larger the alpha, reaching statistically significant relationships with increases of variance accounted for of 13%, 25%, 36%, and 14%, respectively, once the influence of the other predictors was controlled. On the other hand, the mean of subscale scores showed a negative, statistically significant relationship with alpha coefficients for *Responsibility* and *Unacceptable thoughts* subscales, once the influence of the other predictors was controlled, with increases of variance accounted for of 28% and 12%, respectively. More information is available at https://osf.io/cm2v3/

Estimating the Reliability Induction

From 156 studies that applied the DOCS, 37 induced reliability from previous studies, implying a 23.7% of reliability induction for this scale (see Figure 1). Reliability induction rates were estimated distinguishing between two types (Shields & Caruso, 2004): "by omission", that is, when researchers make no reference to the reliability of test scores, and "by report", that is, when researchers report reliability estimates from previous studies. From the 37 studies that induced reliability, 18 (48.7%) omitted any reference to DOCS scores reliability, whereas the remaining 19 studies (51.3%) induced it from previous studies. In particular, of these 19

Table 3
Results of the subgroup analyses applied on alpha coefficients for DOCS total score, taking categorical moderator variables as independent variables

Variable			95% CI		
	k	α_+	LL	LU	ANOVA results
Test version:					
Original (English)	57	.926	.919	.932	
Italian	5	.931	.907	.949	$F(5, 66)=0.741, p=.595$
Spanish	4	.929	.902	.948	$R^2=0.0$
Turkish	1	.870	.745	.934	$Q_W(66)=610.220$,
Icelandic	2	.918	.868	.949	$p<.0001$
Other	3	.915	.875	.942	
Test version (dich.):					$F(1, 70)=0.148, p=.702$
Original (English)	57	.926	.919	.932	$R^2=0.0$
Other	15	.923	.908	.935	$Q_W(70)=628.481$, $p<.0001$
Administration format:					$F(2, 69)=0.378, p=.687$
Pencil and paper	34	.922	.913	.931	$R^2=0.0$
Online	34	.927	.918	.935	$Q_W(69)=604.725$,
Not clear	4	.930	.903	.949	$p<.0001$
Continent:					
North America	49	.927	.920	.933	$F(4, 67)=0.407, p=.803$
Europe	13	.922	.905	.935	$R^2=0.0$
Oceania	7	.922	.895	.942	$Q_W(67)=622.158$,
Asia	2	.905	.846	.941	$p<.0001$
Central America	1	.920	.851	.957	
Study focus:					$F(1, 70)=0.742, p=.392$
Psychometric	22	.921	.910	.931	$R^2=.001$
Applied	50	.927	.920	.933	$Q_W(70)=623.651$, $p<.0001$
Psychometric focus:					$F(1, 20)=0.211, p=.651$
DOCS	17	.920	.906	.932	$R^2=0.0$
Other	5	.926	.900	.944	$Q_W(20)==152.47$, $p<.0001$
Target population:					
Community	18	.935	.926	.943	$F(4, 67)=6.30, p=.0002$
Undergraduate	20	.921	.911	.930	$R^2=.36$
Subclinical	4	.911	.882	.932	$Q_W(67)==351.82$ $p<.0001$
Clinical	23	.911	.899	.921	
Mixed	7	.947	.934	.957	

Note: K = number of studies; α_+ = mean coefficient alpha; LL and LU = lower and upper 95% confidence limits for α_+; F = Knapp-Hartung's statistic for testing the significance of the moderator variable; Q_W = statistic for testing the model misspecification; R^2 = proportion of variance accounted for by the moderator

Table 4
Results of the multiple meta-regression applied on alpha coefficients for Total subscale, taking as predictors the mean and the percentage of patients diagnosed by OCD (k = 54)

Predictor variable	b_j	t	p	Full model fit
Intercept	2.876	21.19	<.0001	$F(2, 51) = 2.67$; $p = .079$ $R^2 = .10$ $Q_E = 413.35$; $p < .0001$
Mean total score	-0.014	-1.61	.113	
Disorder (%OCD)	0.0004	0.19	.848	

Note: b_j = partial unstandardized regression coefficient of each predictor; t = statistic for testing the significance of the predictor (with 51 degrees of freedom); p = probability level for the t statistic; F = Knapp-Hartung's statistic for testing the significance of the full model; Q_E = statistic for testing the model misspecification; R^2 = proportion of variance accounted for by the predictors; ΔR^2 = increase in R^2 as consequence of including in the model a predictor once the other predictors had already been introduced

studies, 11 (57.9%) induced reliability accurately (i.e., by reporting specific estimates from previous studies), and 8 (42.1%) induced it vaguely (i.e., not reporting specific estimates).

Discussion

We conducted an RG meta-analysis of DOCS scores (Abramowitz et al., 2010), a well-established measurement tool to assess obsessive-compulsive symptomatology, aiming to characterize how reliability of DOCS scores varies from one application to the next. The reliability induction rate of the DOCS was also calculated. Guidelines proposed in the literature recommend internal consistency reliability over .8 for research purposes and over .9 for clinical practice (Charter, 2003; Nunnally & Bernstein, 1994). As regards internal consistency, our results evidenced excellent reliability of the DOCS scores, both for total score and subscales, with average alphas over .9 in all cases. The one exception was the *Contamination* subscale, with average alpha slightly under .9. Therefore, the DOCS is a highly recommended measurement tool in assessing obsessive-compulsive symptomatology both for research and clinical purposes, at least regarding internal consistency.

Several RG meta-analyses have been conducted on other scales that assess obsessive-compulsive symptomatology in adults. The MOCI exhibited an average alpha of .76 (Sánchez-Meca et al., 2011), the Y-BOCS an average alpha of .87 (López-Pina et al., 2015), and the PI, the PI-R, and the PI-WSUR average alphas of .93 (Núñez-Núñez et al., 2020; Rubio-Aparicio et al., 2020; Sánchez-Meca et al., 2017). In this RG meta-analysis, the DOCS scores exhibited an average alpha coefficient of .925, clearly larger than that obtained with the MOCI and the Y-BOCS, and practically identical to that of the PI, PI-R, and PI-WSUR. As the PI, PI-R, and PI-WSUR contain 60, 41, and 39 items, respectively, the DOCS is more efficient than the different versions of the Padua Inventory of Obsessions and Compulsions, as it reaches the same reliability with only 20 items.

As expected, the DOCS alpha coefficients exhibited large heterogeneity. A multiple meta-regression model applied on alpha coefficients for the DOCS total score did not reach relevant predictive power; therefore, we were unable to identify study characteristics that explain alpha coefficient variability. As for subscales, simple meta-regression models applied on alpha coefficients of the scores of each subscale enabled us to identify several study characteristics that systematically exhibited a relevant association to the alpha coefficients. As psychometric theory states, the larger the variability of the test scores (e.g., standard deviation), the larger the reliability (cf., e.g., Botella et al., 2010). A positive and relevant association with alphas was also obtained for the mean subscale score, the mean and SD of age, and both percentage of males and percentage of participants with OCD in the sample. In addition, alphas of subscale scores were generally larger for clinical samples and lower for samples with university undergraduates. However, when multiple meta-regression models were applied to each subscale, the only study characteristic exhibiting a statistically significant contribution to the model was the standard deviation of the subscale scores. In addition, the mean subscale scores also exhibited a relevant contribution to the model in the *Responsibility* and *Unacceptable Thoughts* subscales. The absence of statistical significance of most moderators in the multiple meta-regression models was due to high collinearity among moderators.

We also intended to estimate the reliability induction rate of the DOCS, that is, the extent to which studies that applied the DOCS committed the malpractice of inducing reliability from other applications of the test. The reliability induction found for the DOCS was 23.7%, a clearly lower rate than that usually found in many other RG meta-analyses. A systematic review of 100 RG meta-analyses on 123 psychological tests and a total of 41,824 primary studies revealed an average reliability induction rate of 78.6% (Sánchez-Meca et al., 2015, July). Similar induction rates were found by Green et al. (2011) in a representative sample of studies published in *Psychological Assessment*. The low reliability induction rate found with the DOCS could be due to the test being quite recent, therefore researchers are now more aware of the need to report reliability estimates with the data at hand instead of inducing these from previous studies (Appelbaum et al., 2018; Heldref Foundation, 1997; Thompson, 1994; Vacha-Haase et al., 2000; Wilkinson & APA TFSI, 1999). In this line, a positive relationship was found by Sánchez-Meca et al. (2015, July) between the reliability reporting rate and the publication year of 100 RG meta-analyses.

Limitations

Although this RG meta-analysis included a large number of studies reporting reliability estimates with the data at hand, missing data on potential moderator variables were common. This tended to limit generalizability of results obtained from the explanatory models applied. In addition, it is worth noting that this RG meta-analysis was mainly based on Cronbach's alpha coefficients. Alpha coefficient has received much criticism as an internal consistency estimate as it is difficult for the restrictive assumptions of the tau-equivalent model to be fulfilled by test scores (McNeish, 2018; Sijtsma, 2009; Yang & Green, 2011). Instead of alpha coefficient, other reliability coefficients, such as omega coefficients, based on the congeneric model are more realistic (Lenz et al., 2020; Watkins, 2017). Furthermore, correlated errors bias the reliability estimate for both alpha and omega coefficients (Gu et al., 2013). In the extent that studies share their databases, using item-item correlation matrices will enable to apply measurement-based meta-analytic structural equation models (Scherer & Teo, 2020). Finally,

unidimensionality is an important assumption for the estimation of internal consistency reliability (McNeish, 2018; Watkins, 2017). The DOCS global scale does not fulfil this requirement, which may explain the discrepancies found between the moderators related to the reliability coefficients for the global scale and the subscales.

As studies report other alternative reliability coefficients to Cronbach's alpha to assess internal consistency, future RG meta-analyses should base their analyses on new coefficients, as omega.

Implications for Future Research and Clinical Practice

As reliability is not an inherent property to the test, but to the test scores obtained in a given test application, researchers should be aware of the need to report reliability estimates with the data at hand. This is in line with recommendations from scientific journals and reporting guidelines proposed by different international organizations, such as the American Psychological Association, the American Educational Research Association, or the National Research Council on Measurement in Education.

Future RG meta-analyses should improve reporting practices, as there is evidence of poor practice in many published RG meta-analyses. A systematic review of 150 RG meta-analyses conducted on psychological tests showed poor compliance with good reporting practices (Sánchez-Meca et al., 2019, July). Therefore, it is advisable to apply any checklist on reporting practices. We recommend REGEMA checklist in particular, for being specifically devised to improve reporting practices of RG meta-analyses (Sánchez-Meca et al., 2021 May).

Regarding clinical practice, the DOCS is seen to have excellent internal consistency reliability for use in clinical purposes, as average alphas of the total scale and subscales were over .9 (Nunnally & Bernstein, 1994). The only exception was the *Contamination* subscale, which presented an average alpha slightly under the cut-off point of .9 (α_+ = .881). On the other hand, the administration format of the DOCS did not affect reliability coefficients; therefore, this test could be applied online instead of face to face, increasing its accessibility.

The DOCS exhibit excellent internal consistency reliability, even larger than other scales that assess obsessive-compulsive symptomatology and which contain more items. This means DOCS is a very useful measurement tool for both research and clinical purposes to assess obsessive-compulsive symptomatology of people with OCD and other related disorders. However, reliability of DOCS scores varies from one application to another, therefore researchers must report reliability estimates with the data at hand. In explaining reliability estimate variability, the standard deviation of the DOCS subscale scores is the most relevant moderator variable.

Acknowledgments

This research was funded with a grant from the Spanish Government, Ministerio de Economía, Industria y Competitividad, and FEDER funds (project n. PSI2016-77676-P).

References

Abramowitz, J.S., Deacon, B.J., Olatunji, B.O., Wheaton, M.G., Berman, N.C., Losardo, D., Timpano, K.R., McGrath, P.B., Riemann, B.C., Adams, T., Björgvinsson, T., Storch, E.A., & Hale, L.R. (2010). Assessment of obsessive-compulsive symptom dimensions: Development and evaluation of the Dimensional Obsessive-Compulsive Scale. *Psychological Assessment*, *22*(1), 180-198. https://doi.org/10.1037/a0018260

Appelbaum, M., Cooper, H., Kline, R.B., Mayo-Wilson, E., Nezu, A.M., & Rao, S.M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*, 3-25. https://doi.org/10.1037/amp0000191

Bonett, D.G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statistics in Medicine, 21*(9), 1331-1335.

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15*(4), 386-397. https://doi.org/10.1037/a0019626

Charter, R.A. (2003). A breakdown of reliability coefficients by test type and reliability methods, and the clinical implications of low reliability. *The Journal of General Psychology, 130*, 290-304. https://doi.org/10.1080/00221300309601160

Cooper, H., Hedges, L.V., & Valentine, J.C. (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). Russell Sage Foundation.

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

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. https://doi.apa.org/doi/10.1037/a0023089

Gu, F., Little, T.D., & Kingston, N.M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau-equivalence and uncorrelated errors are violated. *Methodology*, *9*(1), 30-40. https://doi.org/10.1027/1614-2241/a000052

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in the meta-analysis with normally distributed responses. *Statistics in Medicine, 20*, 1771-1782. https://doi.org/10.1002/sim.791

Heldref Foundation (1997). Guidelines for contributors. *Journal of Experimental Education, 65*, 95-96.

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. https://doi.org/10.1080/07481756.2002.12069054

Higgins, J.P.T., & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558. https://doi.org/10.1002/sim.1186

Huedo-Medina, T.B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods*, *11*(2), 193-206. https://doi.org/10.1037/1082-989x.11.2.193

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*, 2693-2710. https://doi.org/10.1002/sim.1482

Lenz, A.S., Ho, C.-M., Rocha, L., & Aras, Y. (2020). Reliability generalization of scores on the Post-Traumatic Growth Inventory. *Measurement and Evaluation in Counseling and Development*. Advance online publication. https://doi.org/10.1080/07481756.2020.1747940

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*(5), 443-469. https://doi.org/10.3102/1076998612466142

López-Pina, J.A., Sánchez-Meca, J., López-López, J.A., Marín-Martínez, F., Núñez-Núñez, R.M., Rosa-Alcázar, A.I., Gómez-Conesa, A., & Ferrer-Requena, J. (2015). The Yale-Brown Obsessive Compulsive Scale: A reliability generalization meta-analysis. *Assessment, 22*(5), 619-628. https://doi.org/10.1177/1073191114551954

Mataix-Cols, D., Rosario-Campos, M.C., & Leckman, J.F. (2005). A multidimensional model of obsessive-compulsive disorder. *The American Journal of Psychiatry, 162*(2), 228-238. https://doi.org/10.1176/appi.ajp.162.2.228

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*, 412-412. https://doi.org/10.1037/met0000144

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

Ñúñez-Ñúñez, R.M., Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., López-Pina, J.A., & López-López, J.A. (2020). *A reliability generalization meta-analysis of the Padua Inventory-Revised (PIR-R)* [Manuscript submitted for publication]. Meta-Analysis Unit, Department of Basic Psychology and Methodology, University of Murcia.

Rodríguez, M.C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322. https://doi.org/10.1037/1082-989X.11.3.306

Rubio-Aparicio, M., Núñez-Núñez, R.M., Sánchez-Meca, J., López-Pina, J.A., Marín-Martínez, F., & López-López, J.A. (2020). The Padua Inventory–Washington State University Revision of obsessions and compulsions: A reliability generalization meta-analysis. *Journal of Personality Assessment*, *102*(1), 113-123. https://doi.org/10.1080/00223891.2018.1483378

Ruscio, A.M., Stein, D.J., Chiu, W.T., & Kessler, R.C. (2010). The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Molecular Psychiatry, 15*, 53-63. https://doi.org/10.1038/mp.2008.94

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*(1), 31-48. https://doi.org/10.1037/1082-989X.13.1.31

Sánchez-Meca, J., López-Pina, J.A., López-López, J.A., Marín-Martínez, F., Rosa-Alcázar, A.I., & Gómez-Conesa, A. (2011). The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis. *International Journal of Clinical and Health Psychology, 11*(3), 473-493.

Sánchez-Meca, J., López-López, J.A., & López-Pina, J.A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *The British Journal of Mathematical and Statistical Psychology, 66*(3), 402-425. https://doi.org/10.1111/j.2044-8317.2012.02057.x

Sánchez-Meca, J., López-Pina, J.A., Rubio-Aparicio, M., Marín-Martínez, F., Núñez-Núñez, R.M., López-García, J.J., & López-López, J.A. (2019, May 27-31). *REGEMA: Guidelines for conducting and reporting reliability generalization meta-analyses* [Paper presentation]. Research Synthesis, Dubrovnik, Croatia. http://dx.doi.org/10.23668/psycharchives.2449

Sánchez-Meca, J., Marín-Martínez, F., Núñez-Núñez, R.M., Rubio-Aparicio, M., López-López, J.A., Blázquez-Rincón, D.M., López-Ibáñez, C., López-Nicolás, R., López-Pina, J.A., & López-García, J.J. (2019, July 8-10). *Reporting practices of reliability generalization meta-analysis: An assessment with the REGEMA checklist* [Paper presentation]. XVI Congress of Methodology of the Social and Health Sciences, Madrid, Spain.

Sánchez-Meca, J., Rubio-Aparicio, M., López-Pina, J.A., Núñez-Núñez, R.M., & Marín-Martínez, F. (2015, July). *The phenomenon of reliability induction in the social and health sciences* [Paper presentation]. XIV Congress of Methodology of Social and Health Sciences, Palma de Mallorca, Spain.

Sánchez-Meca, J., Rubio-Aparicio, M., Núñez-Núñez, R.M., López-Pina, J.A., Marín-Martínez, F., & López-López, J.A. (2017). A reliability generalization meta-analysis of the Padua Inventory of Obsessions and Compulsions. *The Spanish Journal of Psychology, 20*, 1-15. https://doi.org/10.1017/sjp.2017.65

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods, 25*(6), 747-775. http://dx.doi.org/10.1037/met0000261

Shields, A.L., & Caruso, J.C. (2004). A reliability induction and reliability generalization study of the cage questionnaire. *Educational and Psychological Measurement, 64*(2), 254-270. https://doi.org/10.1177/0013164403261814

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107-120. https://doi.org/10.1007/s11336-008-9101-0

Streiner, D.L., & Norman, G.R. (2008). *Health measurement scales: A practical guide to their development and use* (4th ed.). Oxford University Press.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20. https://doi.org/10.1177/0013164498058001002

Vacha-Haase, T., Kogan, L.R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those of test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509-522. https://doi.org/10.1177/00131640021970682

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metaphor package. *Journal of Statistical Software, 36*, 1-48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., López-López, J.A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods, 20*(3), 360-374. https://doi.org/10.1037/met0000023

Watkins, M.W. (2017). The reliability of multidimensional neuropsychological measures: From alpha to omega. *The Clinical Neuropsychologist*, *31*(6-7), 1113-1126. http://dx.doi.org/10.1080/13854046.2017.1317364

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

Yang, Y., & Green, S.B. (2011). Coefficient alpha: A reliability coefficient for the 21st Century? *Journal of Psychoeducational Assessment*, *29*, 377-392. https://doi.org/10.1177%2F0734282911406668