Reliability Generalization Meta-Analysis of the Padua Inventory-Revised (PI-R)

María Rubio-Aparicio[1], Julio Sánchez-Meca[1], Rosa Mª Núñez-Núñez[2], José Antonio López-Pina[1], Fulgencio Marín-Martínez[1], José Antonio López-López[1,3]

1. University of Murcia (Spain); 2. University Miguel Hernández (Spain); 3. University of Bristol (UK)

This research was funded by a grant from the Ministerio de Economía y Competitividad of the Spanish Government and Fondo Europeo de Desarrollo Regional (FEDER) (Project No. PSI2016-77676-P).

Introduction

- **Obsessive–compulsive disorder (OCD)** is a mental disorder characterized by the presence of obsessions, compulsions, or both.
- In the psychological practice, **several questionnaires** have been developed to evaluate the symptomatology and severity of Obsessive-Compulsive Disorder (OCD).
- **The Padua Inventory (PI)** of Sanavio is one of the measurement instruments most widely used to assess obsessive-compulsive symptoms (Sanavio, 1988).
- A number of shorter **versions of the PI** can also be found in the literature.
- This is the case of the **Padua Inventory Revised (PI-R)** developed by Van Oppen, Hoekstra, and Emmelkamp (1995).

PI-R (Van Oppen, Hoekstra, and Emmelkamp,1995)

- PI-R consists of **41 items** and **five subscales** adapted to Dutch language: Impulses (7 items), Washing (10 items), Checking (7 items), Rumination (11 items) and Precision (6 items).

- **Higher scores** indicate **greater severity** of obsessive–compulsive symptoms.

- The internal consistency values in the OCD sample were .89 for the total scale and.77-.93 for the subscales; In the anxiety sample was .92 of the total scale and .65-.77 in the subscales, and in the community sample of .92 the total scale and .66-.87 in the subscales.

Reliability Generalization (RG)

- **Reliability of psychological tests** depends on the composition and characteristics of the samples of participants and the application context
- **Reliability is not an inherent property of the test** but of the scores in a given application of the test.
- Since reliability varies in each test administration, **meta-analysis is a suitable method to statistically integrate the reliability** estimates obtained across different applications of a test.
- Vacha-Haase (1998) coined the term **reliability generalization** (RG) to refer to this type of meta-analysis.

Objectives

An **RG meta-analysis** of the **empirical studies** that applied the **PI-R** was carried out in order to:

(a) estimate the **average reliability** (for the total scale and subscales)

(b) examine the **variability** among the **reliability estimates**

(c) search for **characteristics of the studies** (moderators) that can be statistically associated to the reliability coefficients.

Method

Selection criteria of the studies

To **be included in the meta-analysis**, each study had to fulfil the following criteria:

- ✓to be an empirical study where the PI-R, or an adaptation maintaining the 41 items, was applied to a sample of at least 10 participants
- ✓to report any reliability estimate based on the study-specific sample
- ✓the paper had to be written in English or Spanish
- ✓samples of participant from any target population were accepted (community, clinical of subclinical populations)
- ✓the paper could be published or unpublished

Method

Searching for the studies

- The **search period** of relevant studies covered from 1988 to September 2017 inclusive.
- The following **databases** were consulted: PROQUEST, PUBMED, and Google Scholar.
- In the electronic searches, the **keywords** “Padua Inventory” were searched throughout the full text of the documents.
- Furthermore, the references of the studies retrieved were also checked in order to identify additional studies that might fulfil the selection criteria.

Method

Data extraction

- mean and standard deviation (SD) of the total scores and the five subscales
- mean and SD of the participants' age
- gender distribution of the sample
- sample ethnicity
- mean and SD of the history of the disorder
- target population
- percentage of clinical participants in the sample;
- type of clinical disorder
- geographical location of the study
- test version (Dutch original vs. other)
- administration format (clinical interview vs. self-report)
- study focus (psychometric vs. applied)
- diagnostic procedure of the participants
- sample size
- time interval (in weeks) for test-retest reliability
- year of the study
- training of the main researcher (psychology, psychiatry, other)

Alpha and **test-retest** coefficients were extracted for the **total scale** and for the **subscales**

Reliability of the coding process was highly satisfactory with kappa coefficients ranging between .96 and 1.0 (mean = .99) and intraclass correlations between .99 and 1.0 (mean = .99)

Method

Reliability estimates

Two types of reliability coefficients

Coefficients alpha to assess internal consistency of the measures → Transformed → Formula proposed by Bonett (2002)

Pearson correlation coefficients to estimate test-retest temporal stability → Transformed → Fisher´s *Z*

To facilitate the interpretation, the results obtained with Bonett´s or Fisher´s *Z* transformations were back-transformed into the original coefficient alpha and Pearson correlation metrics

Method

Statistical analyses

- A **random-effects model** was assumed estimating the between-studies variance by **restricted maximum likelihood**
- The 95% confidence interval around each overall reliability estimate was computed with the **improved method** proposed by Hartung (1999)
- **Heterogeneity** of the reliability coefficients was investigated by constructing a forest plot and by calculating the Q test and the I^2 index.
- **Moderator analyses** were performed through weighted ANOVAs and meta-regression analyses for qualitative and continuous variables, respectively.
- **Mixed-effects models** were assumed, using the **improved method** proposed by Knapp and Hartung to test the statistical significance of moderator variables

All statistical analyses were carried out with the *metafor* package in *R*

Results

Figure 1. Flowchart of the selection process of the studies.

- The search yielded a total of 1,335 references, out of which 1,234 were removed for different reasons.

- The remaining 101 references were empirical studies that had applied the PI-R.

- Out of them, **26** (25.7%) **reported some reliability estimate with the data at hand**, whereas the remaining **75 studies** (74.3%) **induced the reliability** of the PI-R from previous applications of the test:

  - "By omission": 41 studies
  - "By report": 34 studies

Results

- All studies were published and written in English.
- Several studies reported reliability coefficients for two or more different samples, so that the database of our RG study included a total of **29 independent samples.**
- The **total sample size** was *N* = 9,411 participants (min. = 13, max. = 2,976), with mean = 325 participants per sample (Median = 190; SD = 560).
- Regarding the **location of the studies**, three continents were represented: Europe with 21 samples (72.4%), Asia with 5 samples (17.2%), and North America with 3 samples (10.3%).

Results

Total Scale/Subscale	*k*	α_+	95% CI LL	95% CI UL	*Q*	I^2
Total scale	24	.926	.913	.937	445.700**	95.2
Impulses	17	.793	.762	.820	167.918**	91.6
Washing	17	.889	.853	.916	763.189**	98.1
Checking	16	.879	.862	.894	155.812**	90.3
Rumination	17	.870	.845	.890	302.926**	94.7
Precision	16	.727	.678	.768	207.116**	93.7

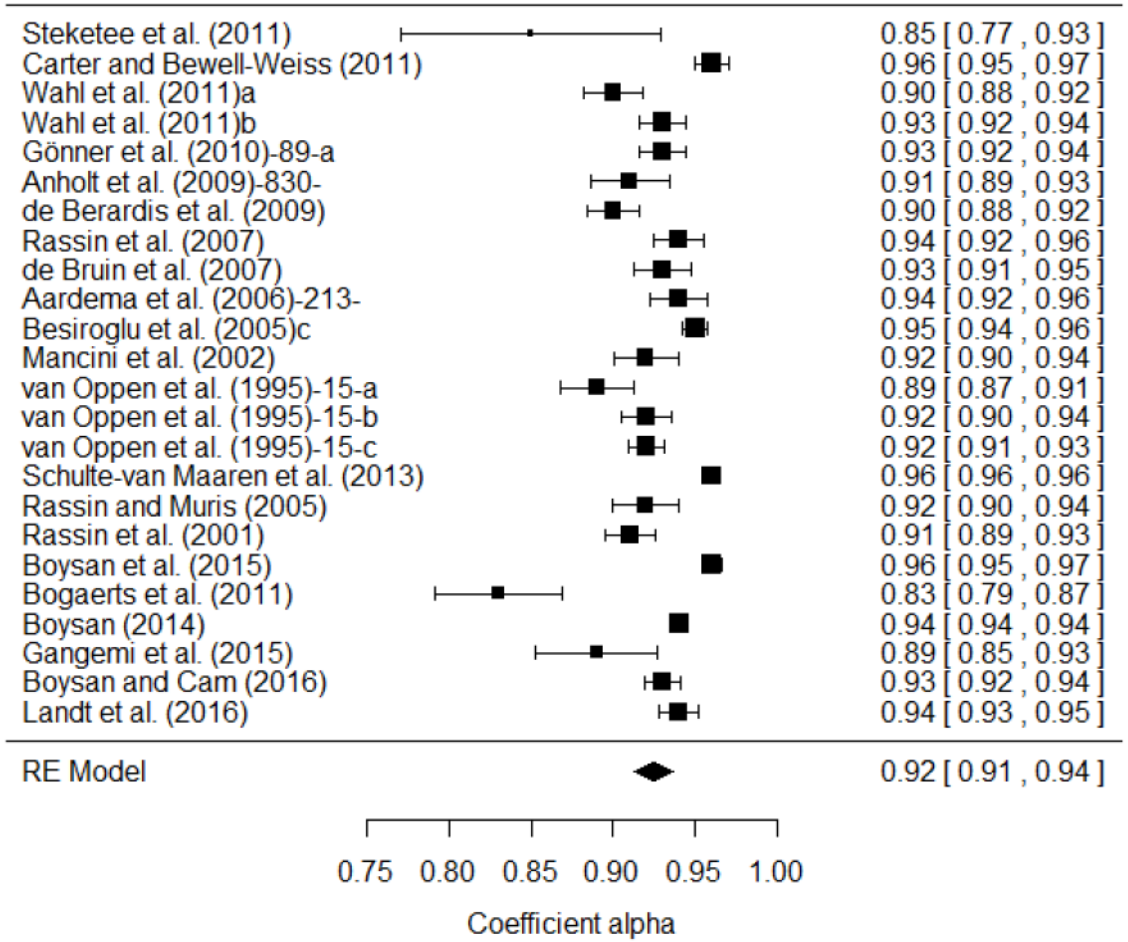

Figure 2. Forest plot displaying the coefficients alpha (and 95% confidence intervals) for the PI-R Total scores.

Regarding the **test-retest reliability**, only two samples reported this kind of reliability for the total score with a **mean** of .910 (95%CI: .879 and .933).

Results

Alpha coefficients presented a **large heterogeneity**, with I^2 indices over 80% in all cases.

The large variability exhibited by the reliability coefficients obtained in different applications of the PI-R was investigated by analyzing the influence of potential **moderator variables.**

Continuous moderator variables

The standard deviation of test scores exhibited a statistically significant relationship with coefficient alpha and with a percentage of variance accounted for of 33%. In particular, this predictor exhibited a positive relationship with coefficients alpha, so that larger coefficients alpha were obtained as the standard deviation of the scores increased.

Results of the simple meta-regressions

Predictor variable	*k*	b_j	*F*	*p*	Q_E	R^2
Mean Total score	24	-0.0003	0.01	.945	430.906***	0.0
SD of Total score	24	-0.0362	9.17	.006	190.177***	.33
Mean age (years)	24	0.0077	0.63	.435	442.652***	0.0
SD of age (years)	24	-0.0039	0.17	.684	436.848***	0.0
Gender (% male)	22	0.0008	1.15	.296	322.678***	.01
% of clinical sample	24	0.0009	0.24	.628	444.805***	0.0
Year of the study	24	-0.0133	1.27	.272	357.773***	.03

Results

Qualitative moderator variables

Results of the weighted ANOVAs

- **Statistically significant differences** were found when comparing the mean coefficients alpha grouped by the **test version** (p = .034), with a 36% of variance of variance explained, the mean reliability being larger for Turkish studies.
- No statistically significant differences were found when comparing coefficients alpha grouped by the **continent** (p = .135), although this moderator explained 12% of the variance among the coefficients. Concretely, the studies conducted in Asia exhibited the largest average coefficient alpha (mean = .949), whereas the lower averages were yielded by studies carried out in Europe and North America (means = .920, .929, respectively).
- It is worth noting that these two moderator variables (test version and continent) are closely related.

Variable	k	α_+	95% CI LL	95% CI LU	ANOVA results
Test version:					
Original (Dutch)	8	.934	.917	.949	$F(5,18) = 3.11$, $p = .034$
German	3	.921	.886	.945	$R^2 = .36$
Italian	3	.905	.861	.935	$Q_W(18)=223.78$, $p<.0001$
Turkish	4	.946	.927	.960	
English	5	.919	.891	.939	
Belgian	1	.830	.675	.911	
Test version (dich.):					$F(1,22) = 1.08$, $p = .311$
Original (Dutch)	8	.934	.913	.949	$R^2 = 0.0$
Other	16	.922	.905	.935	$Q_W(22)=395.88$, $p<.0001$
Study focus:					$F(1,22) = 0.04$, $p = .839$
Psychometric	8	.927	.904	.945	$R^2 = 0.0$
Applied	16	.925	.908	.938	$Q_W(22)=436.21$, $p<.0001$
Psychometric focus:					$F(1,6) = 0.89$, $p = .381$
PI-R	6	.923	.891	.945	$R^2 = 0.0$
Other	2	.940	.893	.967	$Q_W(6)=99.07$, $p<.0001$
Continent:					
Europe	18	.920	.905	.933	$F(2,21) = 2.20$, $p = .135$
N. America	2	.929	.876	.960	$R^2 = .12$
Asia	4	.946	.923	.962	$Q_W(21)=411.23$, $p<.0001$
Target population:					
Community	4	.928	.891	.952	$F(3,20) = 0.19$, $p = .901$
Undergraduate	8	.923	.897	.942	$R^2 = 0.0$
Clinical	7	.922	.893	.943	$Q_W(20)=318.96$, $p<.0001$
Main researcher:					$F(2,21) = 0.42$, $p = .662$
Psychologist	15	.929	.912	.942	$R^2 = 0.0$
Psychiatrist	8	.923	.898	.942	$Q_W(21)=404.01$, $p<.0001$
Both	1	.900	.782	.954	

Conclusions

- **Several guidelines** have been proposed in the psychometric literature to assess the adequacy and relevance of reliability coefficients.
- In general, it is accepted that coefficients alpha must be over .70 for **exploratory research**, over .80 for **general research** purposes, and over .90 when the test is used for taking **clinical decisions** (Nunnally & Bernstein, 1994).
- Based on these guidelines, our findings demonstrated the **good reliability of the PI-R** total scores, both for **screening and clinical purposes**.
- The results also demonstrate how **reliability depends on** the application **context** and the composition and variability of the **samples**.
- In particular, as expected from psychometric theory, a strong positive relationship was found with the **standard deviation of test scores**.
- Another characteristics of the studies that exhibited a statistical relationship with coefficients alpha was **the test version**.

Thanks!

Reliability Generalization Meta-Analysis of the Padua Inventory-Revised (PI-R)

María Rubio-Aparicio[1], Julio Sánchez-Meca[1], Rosa Mª Núñez-Núñez[2], José Antonio López-Pina[1], Fulgencio Marín-Martínez[1], José Antonio López-López[1,3]

1. University of Murcia (Spain); 2. University Miguel Hernández (Spain); 3. University of Bristol (UK)

This research was funded by a grant from the Ministerio de Economía y Competitividad of the Spanish Government and Fondo Europeo de Desarrollo Regional (FEDER) (Project No. PSI2016-77676-P).