Review

ijbd

International Journal of Behavioral Development
1–13
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/01650254221128280
journals.sagepub.com/home/jbd
SAGE

The Adult Prosocialness Behavior Scale: A reliability generalization meta-analysis

Laura Badenes-Ribera[1], Carmen Duro-García[1], Carmen López-Ibáñez[2], Manuel Martí-Vilar[1] and Julio Sánchez-Meca[2]

Abstract

The Adult Prosocialness Behavior Scale (APBS) is most often used to measure adult prosociality. We conducted a reliability generalization meta-analysis to compute the average APBS reliability and examine the heterogeneity among reliability estimations and the influence of moderator variables. An exhaustive search identified 74 articles that applied the APBS with 16 items assessed on a 5-point Likert-type scale. Of these, 58 had reliability coefficients with the current data, and 76 reliability estimates were provided. Random- and mixed-effects models were used. The average reliability coefficient was .903 for Cronbach's alpha, .896 for McDonald's omega, and .674 for test–retest. Moderator analyses were used to create a predictive model in which the target population and study language accounted for 48.7% of the total variability among Cronbach's alpha coefficients. Although the APBS has shown satisfactory internal consistency, it can vary as a function of several factors.

Keywords

Meta-analysis, reliability generalization, prosociality, Prosocialness Behavior Scale

Prosociality is defined as a "set of voluntary actions one may adopt to help, take care of, assist, or comfort others" (Caprara et al., 2005, p. 77). The Adult Prosocialness Behavior Scale (APBS, Caprara et al., 2005) is the only specific prosociality scale developed for adults, aimed at understanding overall prosociality. It specifically assesses individual tendencies to perform a series of prosocial behaviors in different contexts for different reasons. The APBS was created in the Italian context and designed to be a one-dimensional scale consisting of 16 items on a 5-point Likert-type scale to identify behaviors and sensations in one of the four areas: sharing, helping, care-taking, and feeling empathy for others' needs and requirements. The behaviors of sharing, helping, and taking care of have typically characterized the assessment of childhood or teenagers' prosocial behavior, while feeling empathy for others and their requests or needs is an addition to the assessment of prosocialness relative to the APBS. As Caprara et al. (2005) point out, in the adult population, one's empathic predispositions or motives are an integral part of the tendency to act prosocially rather than merely a correlate of such tendencies. The APBS is a brief scale that is easy to use and understand and is a valuable tool for researchers. It has been translated into and adapted to a variety of languages and cultures, such as Chile (Mieres-Chacaltana et al., 2020) and Argentina (Regner & Vignale, 2008). As mentioned above, the APBS was devised as a one-dimensional scale (Caprara et al., 2005); however, some studies have suggested three factors (i.e., helping, empathy, and sharing; Biagioli et al., 2016) or two factors (i.e., prosocial behavior, and empathy and emotional support; Carrizales et al., 2019; Rodriguez et al., 2017). Nevertheless, most studies assume only one factor, as in this study, and provide reliability estimates for the entire scale.

Reliability is not an inherent property of the test, but depends on the composition and variability of the sample to which the test is applied. As reliability changes with each test application, a meta-analysis is an ideal method for examining the factors that can affect or explain variability. The term "reliability generalization" (RG) was coined by Vacha-Haase (1998), who understood it as a meta-analytic procedure that statistically integrates the reliability estimates extracted from previously tested studies (e.g., in different samples and different contexts).

An RG meta-analysis can estimate the average reliability of test scores in different studies and situations, examine the degree of heterogeneity among the reliability coefficients of different

[1] University of Valencia, Spain
[2] University of Murcia, Spain

Corresponding author:
Laura Badenes-Ribera, University of Valencia, 46010 Valencia, Spain.
Email: laura.badenes@uv.es

samples in different contexts, and identify the characteristics related statistically to the reliability estimates (Rodriguez & Maeda, 2006; Sánchez-Meca et al., 2013; Vacha-Haase et al., 2002). Even though reliability estimates estimations of APBS scores are generally high, the fact that they are highly heterogeneous shows that there must be study characteristics that affect the magnitude of the reliability estimates of the scores in the test and, therefore, the accuracy of the measurement.

An RG meta-analysis can thus improve the current information on reliability estimates and factors associated with variations in score reliability in different situations in various populations. This information would help researchers and professionals determine the most important characteristics of the studies with regard to the accuracy of the measurement and thus would allow them to administer the APBS in conditions that maximize the likelihood of obtaining high reliability estimates, that is, an accurate measurement of the subjects' prosocial conduct, which would affect the quality of the information obtained by the APBS.

Purpose of the Study

To our knowledge, variability in the reliability of APBS scores through different applications has not yet been investigated. Thus, our aim was to carry out an RG meta-analysis of previous APBS applications: (1) to estimate the overall reliability of APBS scores; (2) to examine reliability estimates variability; (3) to search for substantive and methodological study characteristics that can be statistically associated with the reliability coefficients; (4) to calculate the APBS reliability induction rate; and (5) to study the generalizability of the results of our RG meta-analysis by comparing the sample characteristics of studies that induced reliability with those that provided reliability estimates from the available data.

Method

An RG meta-analysis was carried out according to the recommendations for conducting and reporting Reliability Generalization Meta-Analyses (REGEMA Checklist; Sánchez-Meca et al., 2021, see Supplemental Material, Table S1).

Inclusion and Exclusion Criteria

For inclusion in the RG meta-analysis, the studies had to fulfill the following inclusion criteria: (1) be an original, quantitative investigation published in a peer-reviewed journal; (2) apply the APBS or any of its adaptations while maintaining the 16-item structure and the 5-point Likert-type scale; (3) report on any reliability estimates based on the study-specific sample for total scores; (4) use of any type of target population; and (5) written in English, Spanish, French, Portuguese, Italian, or German. No limits were set on study dates, geographical locations, or subject ages. Although the APBS was devised to be applied to adults, it has also been applied to children and adolescents. Therefore, we did not restrict the subjects' age range and included non-adult samples. The exclusion criteria were as follows: $N=1$ or case series studies, and those that used another version without the 16-item APBS structure. The selection criteria were the same for the studies that induced reliability (i.e., the reported reliability was not based on the study-specific sample), with the exception of (3), which was also analyzed to compare the different sample characteristics of studies that reported reliability with those that induced it. It should be noted that these studies were not included in the meta-analytic calculations but were analyzed to determine the extent to which both study types used samples from participants with similar sociodemographic characteristics. If inducing and reporting studies exhibit similar sociodemographic characteristics, then the meta-analytic results can be valid for any studies that applied the APBS, regardless of whether they induced reliability.

Search Strategy

Online searches were carried out in February 2021 in the Web of Science, Medline (via PubMed), Scopus, PsycInfo, Science Direct, ProQuest, PubPsych, Psicodoc, and PsycArticles databases, using the following terms: "Prosocial Behavior" and "Caprara," "Prosocial Behavior" and "Scale," and "Prosocialness Scale." Articles in the database that cited the studies by Caprara et al. (2005) were also assessed. References in the collected studies were analyzed to identify those that satisfied the selection criteria. On 6 June, a final search was performed using Google Scholar and other databases to locate all the articles that cited the study by Caprara et al. (2005).

Data Extraction

Data extraction was performed by two independent coders, who extracted the characteristics of the studies in a standardized and systematic manner by applying the established protocol. Interrater reliability was acceptable; the mean intraclass correlation was .999 ($SD=0.002$), ranging from .992 to 1 for continuous variables, and a mean kappa coefficient of .963 ($SD=0.053$), from .883 to 1 for categorical variables. Disagreements were resolved by consensus.

The following characteristics were extracted from the studies: (1) geographical location (country), (2) research design (longitudinal vs. cross-sectional), (3) purpose of the study (psychometric vs. applied), (4) sampling method (convenience vs. randomized sample), (5) sample size, (6) setting (target population), (7) mean and standard deviation of the participants' age, (8) gender distribution (percentage of males), (9) ethnicity (percentage of Caucasian), (10) test version (original version vs. other), (11) total test score mean and standard deviation, (12) year of publication, (13) study language, (14) the main researcher's qualifications (psychology vs. other), and (15) whether funding was obtained (yes/no). When the test–retest reliability was reported, the time interval between the two measures was recorded.

Reliability Estimates

The meta-analysis considered three types of reliability coefficients: Cronbach's alpha and McDonald's omega coefficients, which assess the reliability of the internal consistency of the measures, and test–retest reliability coefficients (Pearson correlation coefficients) to assess temporal stability. To normalize the distributions and stabilize the variances, the reliability coefficients were transformed. Cronbach's alpha and McDonald's omega coefficients were transformed by applying Bonett's (2002) formula, and Pearson correlation coefficients were transformed to Fisher's *Z* (Sánchez-Meca et al., 2013). The average reliability coefficients and their confidence limits obtained by Bonett's or Fisher's Z transformations were then back-transformed into Cronbach's alpha and McDonald's omega coefficients and Pearson correlation metrics, respectively, to facilitate the interpretation of the results from each meta-analysis.

Statistical Analysis

Independent two-level meta-analyses were performed for McDonald's Omega coefficients and test–retest reliability (Pearson correlation coefficients) reported in at least five independent samples (Dimitrov, 2002; Sawilowsky, 2000). In addition, a three-level meta-analysis was performed for Cronbach's alpha coefficients, as several studies have reported more than one reliability coefficient for different samples in the same study. The three-level model considers the hierarchical structure of the data and therefore can deal with the dependency problem by including different samples from the same study for an additional analysis level (Beretvas & Pastor, 2003; Konstantopoulos, 2011; Van den Noortgate et al., 2015). The third level factor was "studies," which modeled the dependency of reliability coefficients that belong to the same study.

To compute summary statistics of reliability coefficient random-effects models were applied in all cases, the coefficients were weighted by inverse variance (Borenstein et al., 2009; López-López et al., 2013; Sánchez-Meca et al., 2013). Restricted maximum likelihood was used to estimate between-study variance (τ^2). In addition, given the scarcity of studies on omega and test–retest reliability and the large heterogeneity exhibited for them, a varying-coefficient model (Bonett, 2010) was applied for McDonald's Omega coefficients and test–retest reliability (Pearson correlation coefficients). This model is recommended because it does not rely on the unrealistic assumption of fixed effects (e.g., constant coefficient) models that there is no variation in population effect sizes (Hunter & Schmidt, 2000), while avoiding assumptions about a random selection of studies from a normally distributed superpopulation central to random coefficient models (Hedges & Vevea, 1998).

An average reliability coefficient and 95% confidence interval were computed for the meta-analyses using the improved method proposed by Hartung and Knapp (2001; see also Sánchez-Meca & Marín-Martínez, 2008), while a forest plot was constructed to visually assess the variability among reliability coefficients in the meta-analyses. Cochran's *Q*-statistic and the I^2 index were calculated. A *Q*-statistic with $p < .05$ and $I^2 > 25\%$ indicated heterogeneity among the effect sizes (Higgins et al., 2003; Huedo-Medina et al., 2006), which was also assessed by the between-study standard deviation (τ) and by calculating a 95% prediction interval.

Moderator analyses were carried out to identify the study characteristics statistically associated with the reliability coefficients for meta-analyses with at least 20 reliability estimates. Weighted analyses of variance (ANOVAs) for categorical variables and meta-regressions for continuous variables were applied to a mixed-effects model using the method recommended by Knapp and Hartung (2003; see also López-López et al., 2013; Viechtbauer et al., 2015). Following Raudenbush's (2009) proposal, the proportion of variance explained by the moderator variable was computed (López-López et al., 2014). Model misspecification was evaluated using the Q_W and Q_E statistics for categorical and continuous moderators, respectively.

A visual inspection of the symmetry of funnel plots (Light & Pillemer, 1984; Sterne & Egger, 2001), followed by a three-level extension of Egger's regression test (Egger et al., 1997) were used to determine whether publication bias might threaten the validity of the results of the meta-analyses. Visual inspection provides a general idea of potential publication bias; however, it does not account for the dependent data within studies. The three-level extension of Egger's regression test explores whether there is a statistically significant association between effect sizes and their standard errors using a three-level approach that considers the dependence between reliability estimates (Fernández-Castilla et al., 2021).

To examine the extent to which the meta-analytic results could be used for other studies that applied the APBS, the sample characteristics of the inducing and reporting studies were compared. These studies were compared in terms of their average and *SD* of age, gender distribution, and mean and *SD* of the APBS total score. All statistical analyses were conducted using the metafor package in R (Viechtbauer, 2010).

Results

Selection Process of the Studies

Figure 1 shows the study selection process. The 1,457 references obtained were exported to RefWorks, which was used to scan and eliminate duplicates and screen the titles. The abstracts were then examined, and 166 articles were eligible, of which 97 applied the APBS. Twenty of these were discarded for applying a short version of the APBS (see Supplementary Material), and three more were discarded for applying the APBS using a 7-point Likert-type scale (Anli, 2019; Rao et al., 2021; Ward & King, 2018), leaving 74 articles that applied the APBS with 16 items using a 5-point Likert-type scale. Of these, 58 reported a reliability estimate based on study-specific data and 16 reported induced reliability or did not report reliability coefficients. The RG study finally analyzed 58 articles that reported reliability coefficients with the data at hand (see Supplemental Material). Of the 58 articles included in the RG meta-analysis, five reported several reliability coefficients from different independent samples (Caprara & Steca, 2005; Layous & Nelson-Coffey, 2020; Martínez-Pampliega et al., 2018; Martí-Vilar et al., 2020; Rodriguez et al., 2021). The database of the present RG meta-analysis thus included a total of 67 independent samples, which provided 76 reliability estimates.

Characteristic of the Studies

The total number of participants in the sample was $N=27{,}091$ (minimum=56, maximum=2,574), with a mean of 404 participants per sample (median=315; *SD*=355). The participants' mean age was 25.85 (Median=20.86, *SD*=13.52), the mean male percentage was 38.9% (Median=40.5%, *SD*=12.3%), and the average Caucasian percentage was 71.6% (Median=76.6%, *SD*=28.7%).

Regarding APBS total scores, for all samples, the average of the mean APBS total scores was 57.44 (Median=58.24, *SD*=8.22, *Min.*=24.67, *Max.*=67.04) and the average of the

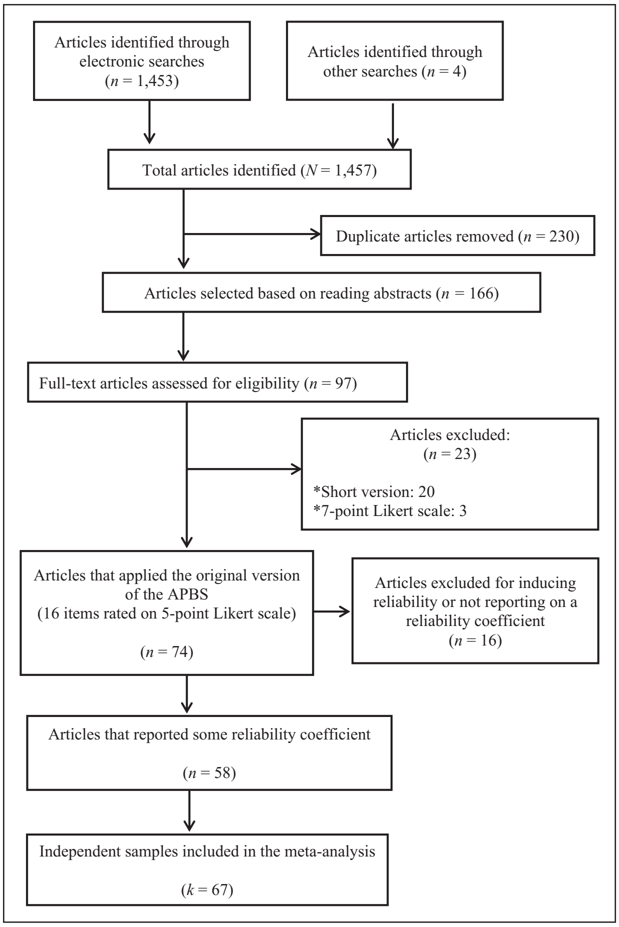
Articles identified through electronic searches (*n* = 1,453)
Articles identified through other searches (*n* = 4)
Total articles identified (*N* = 1,457)
Duplicate articles removed (*n* = 230)
Articles selected based on reading abstracts (*n* = 166)
Full-text articles assessed for eligibility (*n* = 97)
Articles excluded: (*n* = 23)
*Short version: 20
*7-point Likert scale: 3
Articles that applied the original version of the APBS (16 items rated on 5-point Likert scale) (*n* = 74)
Articles excluded for inducing reliability or not reporting on a reliability coefficient (*n* = 16)
Articles that reported some reliability coefficient (*n* = 58)
Independent samples included in the meta-analysis (*k* = 67)

Figure 1. REGEMA Flow Diagram of Study Selecting Process.

standard deviation for APBS total scores was 9.86 (Median= 10.08, *SD*=2.79, *Min.*=2.31, *Max.*=15.68).

Of the 67 independent samples, 57 (85.1%) appeared in articles written in English and the 10 remaining samples (14.9%) were published in Spanish. The studies were conducted in 14 different countries: Italy (22 samples; 32.8%), Spain (14 samples; 20.9%), United States (11 samples; 16.4%), Europe (40 samples; 59.7%), North America (12 samples; 17.9%), South America (11 samples; 16.4%), Asia (3 samples; 4.5%), and Africa (1 sample; 1.5%). All studies used convenience samples, and most of them had cross-sectional research designs (83.6%).

Mean Reliability and Heterogeneity

Independent meta-analyses were performed for each type of reliability coefficient. Several studies have reported separate alpha coefficients for different participant samples. A three-level RG meta-analysis was carried out to take into account the dependence between Cronbach's alpha coefficient values in the same study, while separate two-level RG meta-analyses were conducted to synthesize McDonald's Omega coefficients and test–retest reliability coefficients.

Table 1 shows the mean reliability of each type of coefficient, 95% confidence interval, 95% prediction interval, *Q* test, and I^2 index for the total APBS scores. Cronbach's alpha coefficient was the most frequently used reliability estimate, with an average coefficient of .903 (95% CI [.893, .912], *k*=65), ranging from .790 to .960 (see Figure 2). McDonald's Omega coefficients showed an average reliability of .896 (95% CI [.853, .927], *k*=6), ranging from .845 to .940 (see Figure 3). The test–retest reliability coefficients presented an average coefficient of .672 (95% CI [.553, .764], *k*=5), ranging from .590 to .790 for a time interval between test–retest administrations from 1 day to 208 weeks (*M*=93.63 weeks, *SD*=77.09) (see Figure 4) and an average coefficient of .674 (95% CI [.499, .797], *k*=4), ranging from .590 to .790 for a time interval between test–retest administrations from 52 to 208 weeks (*M*=117 weeks, *SD*=65.43) (see Figure 5). Similar reliability estimates were found when the varying-coefficient meta-analytic model was applied to McDonald's Omega coefficients (ω_+=.896, 95% CI [.859, .922], *k*=6) and the test–retest reliability coefficients (r_+=.672, 95% CI [.654, .687], *k*=5; r_+=.674, 95% CI [.663, .686], *k*=4). In addition, there was evidence that all reliability coefficients were heterogeneous, as all *Q*-statistics were statistically significant, and the I^2 indices ranged from 85.8% to 95.2%.

Analysis of Moderator Variables

Moderator analyses were performed for the reliability coefficients with at least 20 estimates. Of the three types of reliability coefficients analyzed (alpha, omega, and test–retest), only alpha coefficients fulfilled this criterion. Mixed-effects simple meta-regressions were applied to evaluate the influence of continuous moderators on alpha coefficients (see Table 2). Male percentage (*p*=.003, R^2=.16) and year of publication of the study (*p*=.011, R^2=.10) had a statistically significant relationship with alpha coefficients, such that the larger the male percentage, the larger the alpha coefficient, and the most recent studies exhibiting lower alpha coefficients than the oldest. In addition, the participants' mean score was significantly related to the alpha coefficient (*p*=.016, R^2=.14), with a negative regression coefficient indicating that the alphas decreased as the participants' mean scores increased. The participants' mean age was marginally related to the alpha coefficient (*p*=.070, R^2=.03), with a positive regression coefficient indicating that the alphas increased as the average participant age increased.

It is worth noting that variability in APBS scores was not a predictor of consistency coefficients. However, an analysis of the relationships between Cronbach's alpha coefficients and standard deviation of APBS scores by target population revealed a relationship between variability exhibited in the APBS total scores in samples recruited from "other population" and consistency reliability (see Table S2), but not in the rest of the subsamples. Figure 6 displays scatter plots of the relationship between the standard deviation of the APBS total scores and Cronbach's alpha coefficient for the target population. It can be noted that in the subsample recruited from "other population" the variability exhibited in APBS scores showed a negative association with the reliability estimate, indicating that studies with higher variability resulted in lower reliability estimates. Nevertheless, the ANOVA performed on the variance of APBS scores by target population did not find statistically significant differences in variability scores, *F*(3, 43)=1.21, *p*=.317, R^2=.124, $Q_W(43)$=3,268.81, *p*<.001, indicating that the variability of scores exhibited was similar among the different target populations (general population: 104.25, 95%

Table 1. Average Reliability Coefficients, 95% Confidence Intervals, and Heterogeneity Statistics for the Adults' Prosocialness Behavior Scale Total Score.

Reliability estimate	k	N	r_+	95% CI		95% PI		Q	I^2	τ
				LL	UL	LL	UL			
Alpha	65	26,175	.903	.893	.912	.806	.951	1,342.20***	95.2	0.122
Omega	6	1,983	.896	.853	.927	.746	.957	76.45***	93.3	0.320
Test-rest (1 day to 208 weeks)	5	1,667	.672	.553	.764	.359	.849	25.03***	85.9	0.142
Test-rest (52–208 weeks)	4	1,611	.674	.499	.797	.239	.884	25.02***	90.6	0.159

Note. k: number of studies; N: total sample size; r_+: average reliability coefficient; CI: confidence interval for r_+; PI: prediction interval for r_+; Q: Cochran's heterogeneity Q-statistic; Q-statistic has $k - 1$ degrees of freedom; I^2: heterogeneity index; τ: between-studies standard deviation estimated using restricted maximum likelihood; LL and UL = lower and upper limits of the 95% confidence and prediction intervals for r_+.
***$p < .0001$.

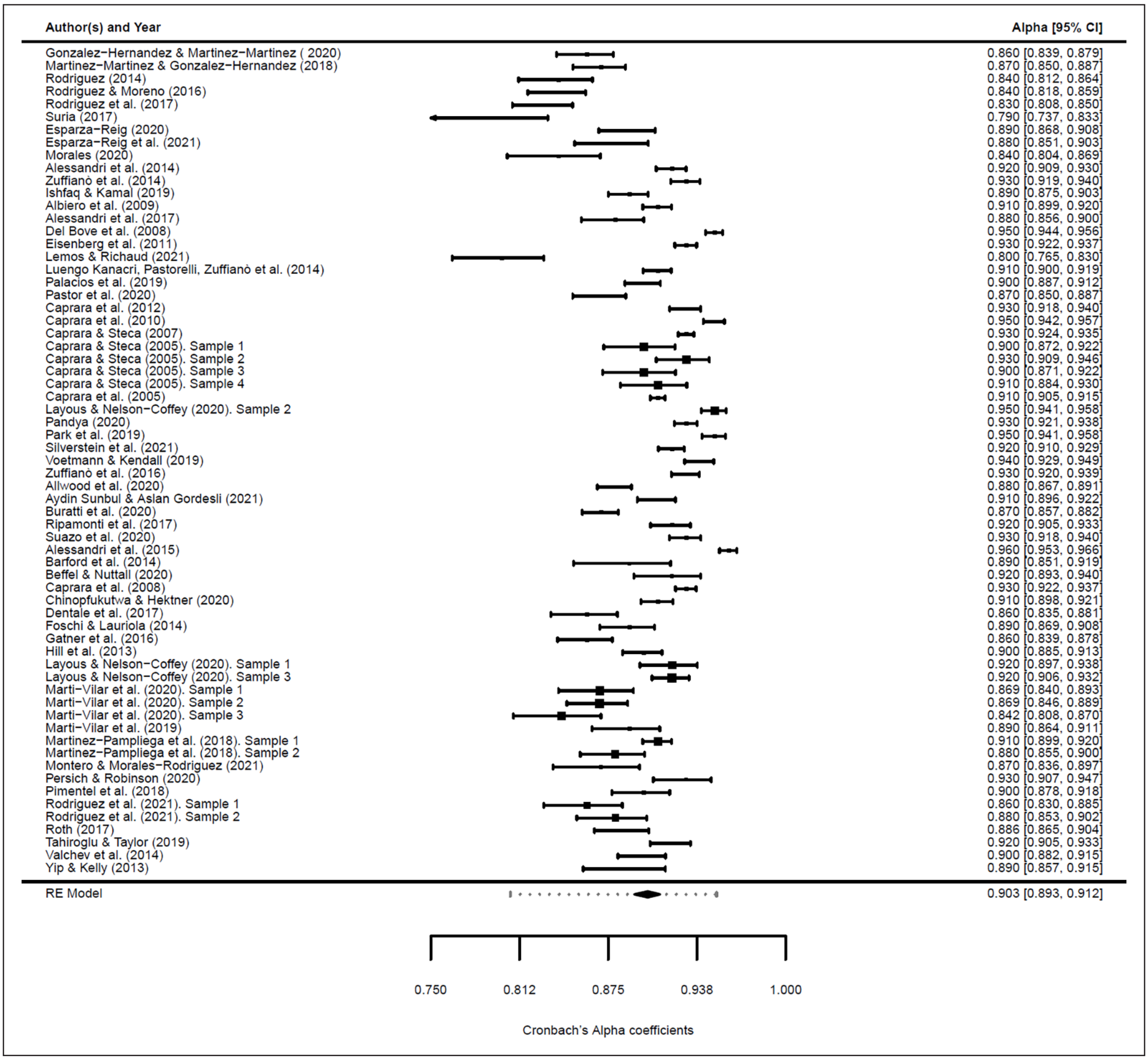

Figure 2. Forest Plot RG MA on Cronbach's Alpha Coefficients.

CI [75.90, 132.61]; other population: 94.71, 95% CI [48.82, 140.60]; school students: 72.75, 95% CI [34.57, 110.94]; and university students: 113.65, 95% CI [91.58, 135.72]).

The influence of categorical moderators on the APBS total score alpha coefficients was analyzed using mixed-effects ANOVAs (see Table 3). Six categorical moderator variables were

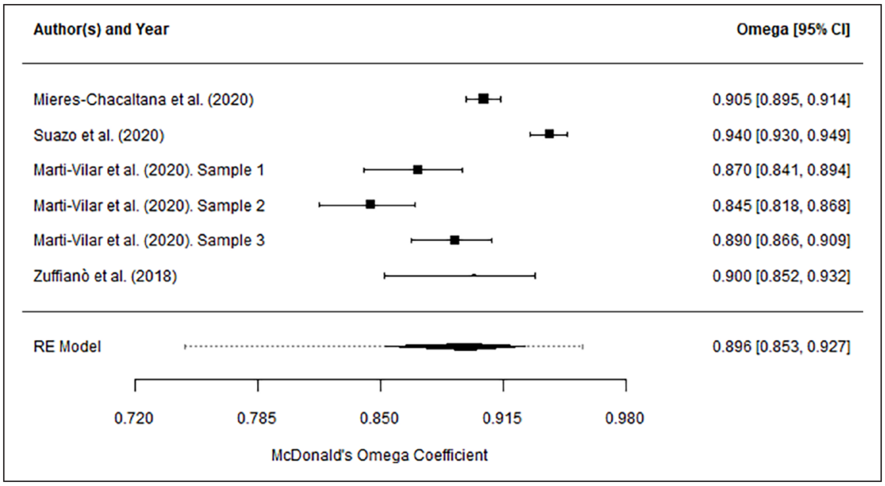

Figure 3. Forest Plot RG MA on McDonald's Omega Coefficients.

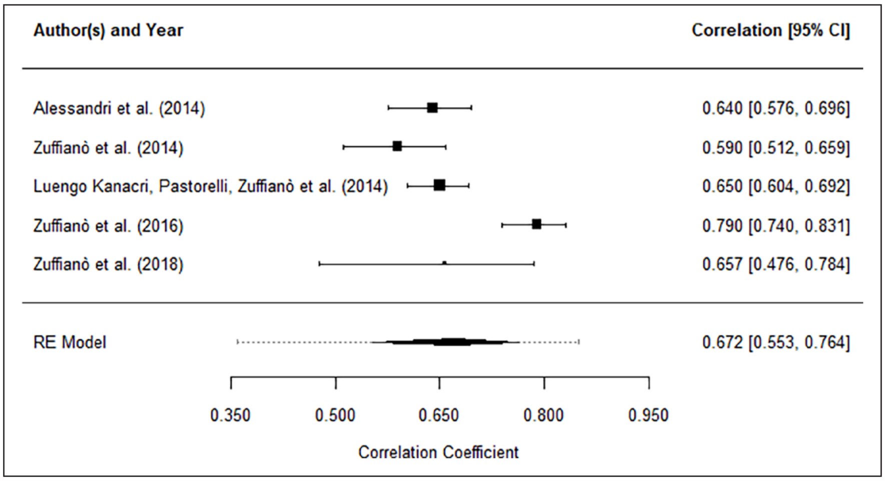

Figure 4. Forest Plot RG MA on Test–Retest Reliability Coefficients (Range: 1 Day to 208 Weeks).

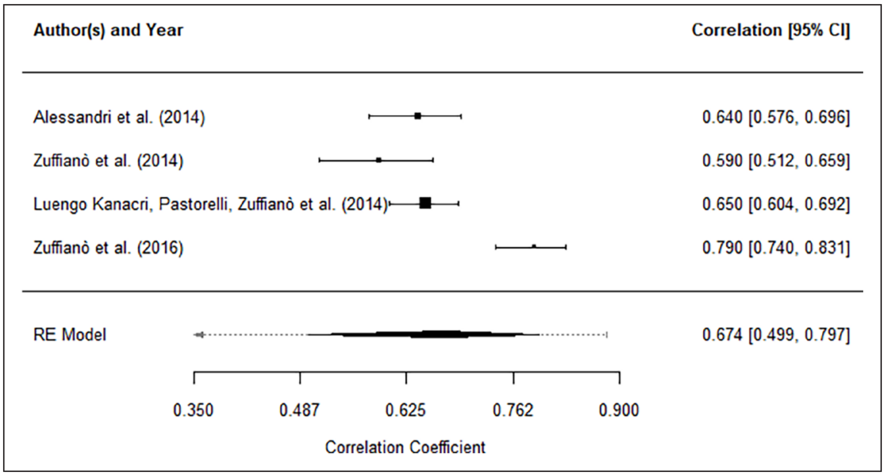

Figure 5. Forest Plot RG MA on Test–Retest Reliability Coefficients (Range: 52–208 Weeks).

Table 2. Results of the Simple Meta-Regressions Applied on Alpha Coefficients for the Adults' Prosocialness Behavior Scale Total Score, Taking Continuous Moderator Variables as Predictors.

Predictor variable	k	b_j	95% CI		F	Q_E	R^2
			LL	UL			
Mean total score	47	−0.207	−0.375	−0.039	5.84*	666.6***	.139
SD total score	47	0.185	−0.380	0.750	.041	787.4***	0
Sample size	65	0.000	−0.000	0.000	1.46	1,323.1***	0
Gender (% male)	63	0.010	0.004	0.017	8.90**	1,055.9***	.156
Mean age (years)	63	0.006	−0.001	0.012	3.27[a]	1,214.6***	.033
SD of age (years)	55	0.010	−0.012	0.032	0.85	1,079.1***	0
Year of the study	65	−0.027	−0.047	−0.006	6.57*	1,165.3***	.097

Note. k: number of studies; b_j: regression coefficient of each predictor; CI: confidence interval; 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); Q_E: statistic for testing the model misspecification; R^2: proportion of variance accounted for by the predictor; LL and UL: lower and upper limits of the 95% confidence interval for b_j.
$*p < .05$. $**p < .01$. $***p < .001$. $^a p = .070$.

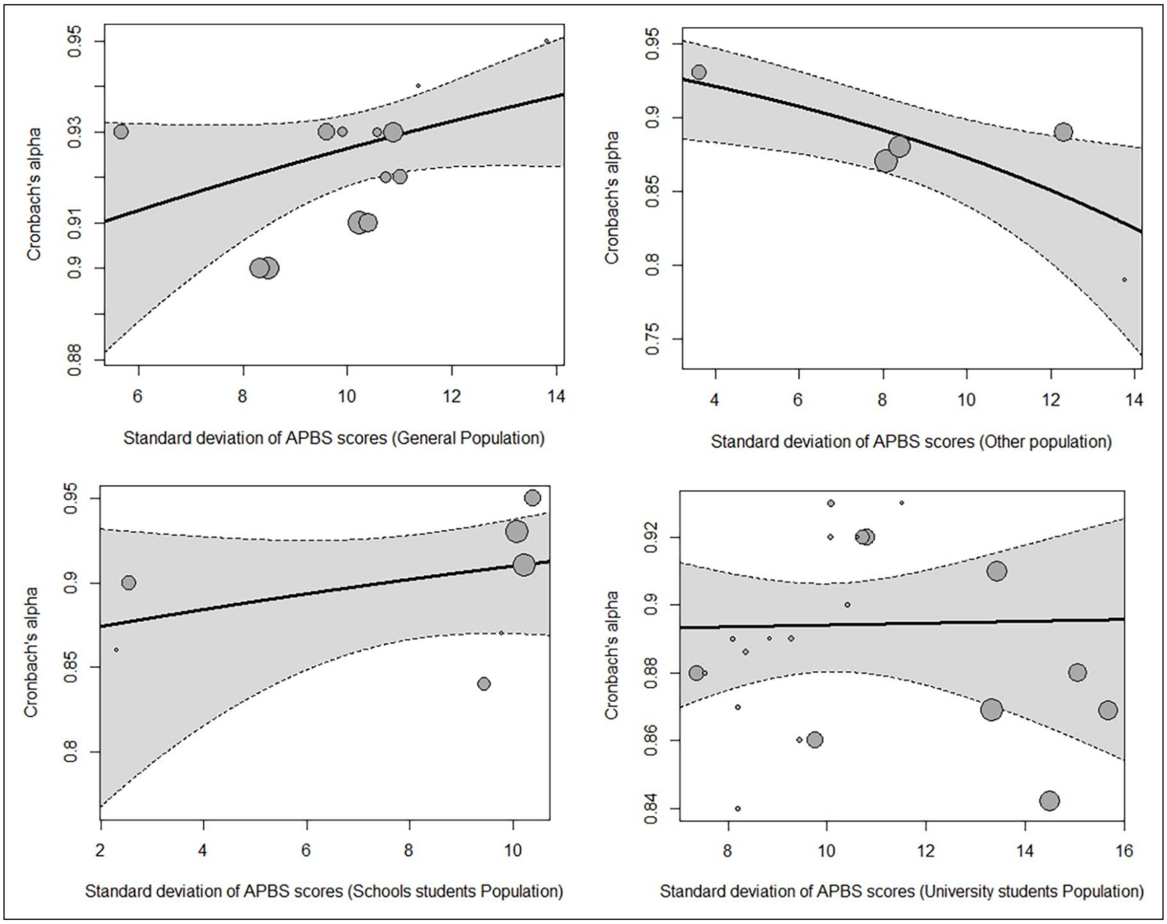

Figure 6. Scatter Plots of the Relationship Between Standard Deviation of APBS Total Scores and Cronbach's Alpha Coefficient by Target Population.

significantly related to alphas: test version, target population, continent, study language, study research design, and financial source. The original (Italian) APBS applications had a larger average alpha coefficient ($\alpha_+ = .920$) than their adaptations to other languages ($\alpha_+ = .893$, $p = .001$, $R^2 = .19$). The target population also had a statistically significant relationship with the alpha coefficients ($p < .001$, $R^2 = .25$), with a larger mean reliability for the studies conducted on samples recruited from the general population ($\alpha_+ = .932$) than those carried out on schoolchildren ($\alpha_+ = .884$). The continent on which the study was carried out also had a statistically significant association with the alpha coefficients ($p = .005$, $R^2 = .19$), with larger mean reliability for studies conducted in Asian countries ($\alpha_+ = .927$), and those carried out in South American countries yielded lower means

Table 3. Results of the Weighted ANOVAs Applied on Alpha Coefficients for the Adults' Prosocialness Behavior Scale Total Score, Taking Categorical Moderator Variables as Independent Variables.

Variable	*k*	*N*	α_+	95% CI		ANOVA results
				LL	LU	
Test version						$F(1, 63) = 10.96$, $p = .001$
Original (Italian)	22	11,462	.920	.908	.931	$R^2 = .188$
Other	43	14,713	.893	.881	.903	$Q_W(63) = 1{,}029.20$, $p < .001$
Study focus						$F(1, 63) = 2.46$, $p = .117$
Psychometric	20	10,046	.908	.897	.917	$R^2 = .028$
Applied	45	16,129	.892	.872	.908	$Q_W(63) = 1{,}308.33$, $p < .001$
Target population						$F(3, 61) = 28.50$, $p < .001$
University students	29	8,020	.897	.884	.909	$R^2 = .252$
General	16	8,542	.932	.921	.942	$Q_W(61) = 1{,}018.83$, $p < .001$
School Students	13	6,280	.884	.863	.902	
Other[a]	7	3,333	.892	.863	.914	
Continent						$F(4, 60) = 15.02$, $p = .005$
Europe	39	17,799	.903	.893	.913	$R^2 = .194$
North America	12	3,156	.918	.899	.933	$Q_W(60) = 999.83$, $p < .001$
South America	10	3,356	.867	.838	.891	
Asia	3	1,433	.927	.895	.949	
Africa	1	431	.900	.812	.947	
Study language						$F(1, 63) = 19.52$, $p < .001$
English	56	23,191	.910	.902	.918	$R^2 = .258$
Spanish	9	2,984	.852	.818	.879	$Q_W(63) = 1{,}006.41$, $p < .001$
Main researcher						$F(1, 63) = 0.60$, $p = .440$
Psychologist	57	23,159	.901	.891	.911	$R^2 = 0$
Other	8	3,016	.911	.886	.931	$Q_W(63) = 1{,}334.56$, $p < .001$
Design research						$F(1, 63) = 5.69$, $p = .017$
Cross-sectional	55	21,396	.898	.887	.907	$R^2 = .082$
Longitudinal	10	4,779	.923	.905	.938	$Q_W(63) = 1{,}230.12$, $p < .001$
Financial source						$F(1, 63) = 4.47$, $p = .034$
Public funding	15	6,957	.919	.902	.932	$R^2 = .063$
No funding	50	19,218	.897	.886	.908	$Q_W(63) = 1{,}295.66$, $p < .001$

Note. ANOVA = analysis of variance; *k*: number of studies; *N*: total sample size; α_+: mean coefficient alpha; CI: confidence interval; 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.

[a]Other: disabilities, psychologists, teachers, and so on.

($\alpha_+ = .867$). The language in which the manuscript was written also revealed a statistically significant relationship with the alpha coefficients ($p < .001$, $R^2 = .26$), with larger means for studies written in English ($\alpha_+ = .910$) than those written in Spanish ($\alpha_+ = .852$). The study's research design was also significantly associated with the alpha coefficients ($p = .017$, $R^2 = .08$), with a larger mean reliability for the studies that used a longitudinal design ($\alpha_+ = .923$) than those that used a cross-sectional design ($\alpha_+ = .898$). Finally, the financial source also showed a statistically significant association with the alpha coefficients ($p = .034$, $R^2 = .06$), with larger mean reliability for the studies that were supported for public funding ($\alpha_+ = .919$) than those that did not receive financial sources ($\alpha_+ = .897$).

Explanatory Models

The Q_E and Q_W statistics (see Tables 2 and 3) suggest that the residual heterogeneity among the reliability estimates was substantial in all models, including a single moderator, even though several moderator variables were significantly associated with the APBS total score alpha coefficients. Thus, a multiple meta-regression was used to identify the most relevant study characteristics to explain the variability of the coefficients. However, given that methodological research recommends at least 10 effect estimates for each predictor (López-López et al., 2014), only five predictors were to include in the model. These predictors were chosen based on the ANOVA results and simple meta-regressions ($R^2 > .10$). Moreover, for the ANOVA results, an additional criterion was that the predictor should be able to be coded as one dummy variable only. As the "continent" variable could not be coded as a dummy variable, it was not included in multiple meta-regression analysis. Five predictors were included in the model: target population, study language, test version, average Adult Prosocialness Behavior Scale (ABPS) total score, and percentage of male participants in the samples. The "target population" (1 = general population, 0 = other population), the "test version" (1 = Original; 0 = Other), and the "language of the study" (1 = English; 0 = Other) were each coded as a dummy variable.

Owing to insufficient data on some variables, $k = 45$ studies were included in this meta-regression (see Table 4). The full model exhibited a statistically significant relationship with

Table 4. Results of the Multiple Meta-Regressions Applied on Alpha Coefficients for the Adults' Prosocialness Behavior Scale Total Scores, Taking as Predictors the Target Population, Language of the Study, Version Test, the Year of the Publication, and the Percentage of Male in the Samples ($k = 45$).

Predictor variable	b_j	*SE*	*z*	*p*
Intercept	2.136	0.320	6.668	<.0001
Target population 1 (General population)	0.206	0.098	2.103	.036
Language of the study (English)	0.388	0.122	3.196	.001
Test version (Original)	−0.012	0.091	−0.127	.899
Percentage of male	0.005	0.004	1.341	.180
Average ABPS total score	−0.007	0.004	−1.578	.115
Global results:	$F(5, 39) = 36.60$, $p < .0001$			
Total τ^2 (intercept-only model): 0.0967	$R^2 = .487$			
Residual τ^2 (full model): 0.0496	$Q_E(39) = 364.97$, $p < .0001$			

Note. b_j: regression coefficient of each predictor; *SE*: standard error for b_j; *z*: statistic for testing the significance of the predictor; *p*: probability level for the z statistic; ABPS: Adult Prosocialness Behavior Scale; *F*: Knapp–Hartung's statistic for testing the significance of the full model; R^2: proportion of variance accounted for by the predictors; Q_E: statistic for testing the model misspecification.

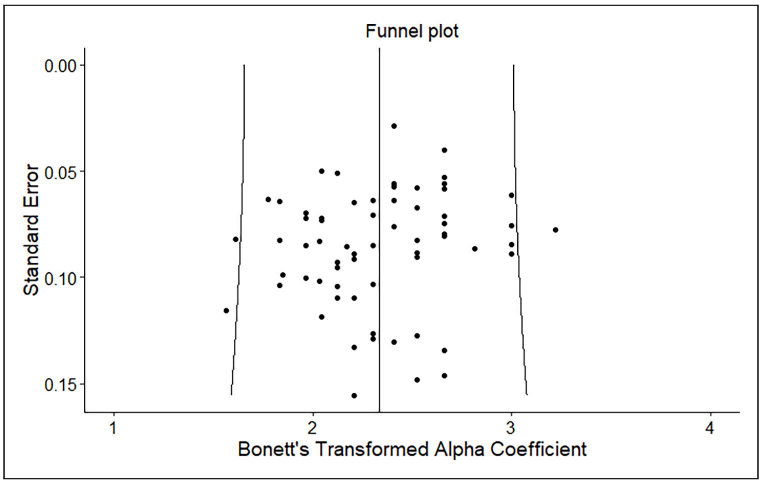

Figure 7. Funnel Plot on Cronbach's Alpha Coefficients to Examine Publication Bias.

alphas, $F(5, 39) = 36.60$, $p < .0001$, with 48.7% of the variance explained. Of the five model predictors, two showed a statistically significant relationship with Cronbach's alpha coefficients after controlling for the influence of the other variables: the study language ($p < .001$) and the target population ($p = .036$); thus, the alpha coefficients obtained in the studies were higher when the studies used samples recruited from the general population and the study was written in English.

Analysis of Publication Bias

Figure 7 displays the funnel plot constructed to examine publication bias in three-level meta-analysis performed on alpha coefficients. Overall, a visual examination of the funnel plots showed that the distributions of the reliability estimates were relatively symmetrical around their means. Although, there were some data points on the lower right and middle left portion plot with no counterparts on the opposite side.

In addition, Egger's test did not reach statistical significance for the meta-analysis of Cronbach's alpha coefficient ($\beta = -1.592$, $SE = 1.642$, $Z = -0.969$, $p = .332$). Therefore, based on the results of these different analyses, publication bias can be reasonably ruled out as a serious threat to the meta-analytic findings.

A Comparison of Studies That Induce and Report Reliability

To determine whether the results can be generalized to studies using the APBS, regardless of whether they reported or induced reliability, the characteristics of the samples in both study types were compared by performing a meta-analysis. The mean and standard deviation of the total score were compared in both types, as well as the mean and standard deviation of the participants' ages and percentage of males.

As previously mentioned, of the 74 articles that had applied the APBS, 58 (78.4%) reported a reliability estimate based on study-specific data, and 16 of these (21.6%) reported either induced reliability or did not find a reliability coefficient. As shown in Table 5, there were no statistically significant differences in any of the sample characteristics between those that induced or reported reliability.

Discussion

The aim of the present RG meta-analysis was to assess the average reliability of the APBS total score and identify the study characteristics that affect the reliability coefficients

Table 5. Results of the Weighted ANOVAs Applied on Different Sample Characteristics of the Studies That Reported and Induced Test Score Reliability.

Variable	*k*	Average	95% CI		ANOVA results
			LL	LU	
Mean Total Score					$F(1, 52) = 0.48, p = .491$
Reported	47	57.43	55.11	59.76	$R^2 = 0$
Induced	7	55.19	49.15	61.24	$Q_W(52) = 48{,}229.37, p < .0001$
Variance Total Score					$F(1, 52) = 0.08, p = .773$
Reported	47	102.58	88.16	116.99	$R^2 = 0$
Induced	7	96.81	58.75	134.88	$Q_W(52) = 7{,}540.24, p < .0001$
Mean age (years)					$F(1, 68) = 1.05, p = .310$
Reported	57	25.77	22.11	29.44	$R^2 = 0.002$
Induced	13	30.14	22.46	37.83	$Q_W(68) = 439{,}052.29, p < .0001$
Variance age (years)					$F(1, 68) = 0.33, p = .568$
Reported	57	30.79	16.08	45.50	$R^2 = 0$
Induced	13	40.63	9.74	71.52	$Q_W(68) = 10{,}778.17, p < .0001$
Male (%)					$F(1, 78) = 0.03, p = .863$
Reported	67	38.84	35.93	41.84	$R^2 = 0$
Induced	15	39.46	33.24	46.04	$Q_W(78) = 1{,}152.44, p < .0001$

Note. ANOVA: analysis of variance; *k*: number of studies; CI: confidence interval; LL and LU: lower and upper 95% confidence limits for average; *F*: Knapp–Hartung's statistic for testing the significance of the moderator variable; R^2: proportion of variance accounted for by the moderator; Q_W: statistic for testing the model misspecification.

obtained by applying the scale. Despite the recommendations of the American Psychological Association (Appelbaum et al., 2018), not all the primary studies that applied the APBS reported a reliability coefficient based on study-specific data: 78.4% of them provided an alpha coefficient and/or McDonald's Omega and/or test–retest coefficients from their own data, representing the different methods of assessing reliability within the Classical Test Theory. Approximately 21.6% of the studies that applied the APBS-induced reliability or did not report a reliability coefficient, which was unexpected, since score reliability needs to be reported when a measure is used. The percentage of articles reporting reliability coefficients with their own data was similar to that of other RG studies (e.g., Rubio-Aparicio et al., 2020).

As mentioned previously, different reliability coefficients have been reported in studies based on specific data. Cronbach's alpha coefficient was the most frequently reported (e.g., Blázquez-Rincón et al., 2022). The average reliability coefficient for total scores, measured as Cronbach's alpha coefficient, was .903, and when assessed as McDonald's Omega coefficient, it was .896. According to the psychometric theory, Cronbach's alpha coefficients over .70 can be considered acceptable for exploratory research. However, in general research, coefficients higher than .80 are recommended and should be higher than .90 in clinical practice (Nunnally & Bernstein, 1994). According to these guidelines, the average internal consistency reliability of the APBS found in this study can be considered adequate for research and acceptable for making clinical decisions. However, prediction intervals for reliability estimates, which indicate the likely range of the future values of reliability estimated in an application of the test in other studies, showed that future internal consistency reliability of the APBS might be adequate for exploratory and general research, but not for clinical practice, given that they might range between .746 and .957.

Regarding test–retest reliability coefficients, Pearson correlations obtained a mean of .672 for the APBS total score. Different opinions exist in the psychometric literature on the guidelines for interpreting the adequacy of test–retest coefficients, given that estimates of test–retest reliability are affected by the period between the test and retest (Charter, 2003; Revelle & Condon, 2018; Watson, 2004). In this RG meta-analysis, the studies used time intervals ranging from 52 to 208 weeks (with the exception of one study that had a time interval of 1 day). A considerable test–retest correlation over a long period indicates temporal stability (Revelle & Condon, 2018; Watson, 2004). The prediction interval for temporal stability reliability estimates of the APBS indicates that future test–retest correlations might range between .259 and .884, which means that the characteristic that is evaluated is not very stable and changes with age, experiences, learning, or the evaluation endeavor.

Moderator analyses provided evidence that nine variables were related to the heterogeneity shown by Cronbach's alpha estimates: mean ABPS total scores, male percentage, year of publication of the study, target population, test version, geographical location (continent), research design, financial source, and study language. It is worth noting that the variability of APBS total scores was not a predictor of the reliability estimates, although the magnitude of Cronbach's alpha coefficients is typically related to the empirical variance of the scores in the samples involved. This lack of significance may have been due to the low range of this explanatory variable (as were 2.31 and 15.68), which was the minimum and maximum total scores standard deviations found in the studies. Nevertheless, an analysis of the relationships between the variances of the observed scores in the different samples (target population) and the corresponding reliability coefficients revealed that the variability of APBS total scores had a negative effect on the reliability estimates in samples recruited from "other populations," suggesting an increase in measurement errors when the test was applied. As Botella et al. (2010) point out, increases in the error variance of measurement when the test is applied imply increases in the empirical variance of the scores and decreases in the reliability estimates. The “other population” variable was composed by a set of samples recruited from people with disabilities, juvenile delinquents and students, psychologists, teachers working in public schools, clergy, subjects were or had been volunteering in hospital settings, and nurses.

Six of the nine variables associated with heterogeneity exhibited by Cronbach’s alpha estimates explained a significant part of it: study language (25.8%), target population (25.2%), continent where the study was conducted (19.4%), the test version applied (18.8%), the percentage of male participants in the samples (15.6%), and the average ABPS total score (13.9%). However, when a multiple regression model was built with five of these six predictors, only two showed a statistically significant relationship with Cronbach’s alpha coefficients after controlling for the influence of the other variables: target population ($p=.036$) and study language ($p=.001$). These predictors accounted for 48.7% of the alpha variance; therefore, larger reliability estimates were noted when the study was carried out with samples recruited from the general population, and the study was written in English. Regarding the study language, it is possible that studies written in Spanish were published in peer-reviewed journals but with a lower impact than those written in English. High-impact journals may require authors to provide higher reliability estimates to be considered for publication, so that the journal’s impact, rather than the study language, may explain this difference among the reliability estimates.

Although the APBS was devised to be applied to adult samples (Caprara et al., 2005), at least 13 studies applied it to teenagers, with a satisfactory average alpha of .884 (Nunnally & Bernstein, 1994). In these studies, the average sample age ranged from 12.3 to 17.5 years, with an average of 15.4 years. Although Caprara and Pastorelli (1993) developed the PBS for children, these studies did not apply the PBS’s children’s but the adults’ version, possibly because Caprara and Pastorelli (1993) validated their version on a sample of children between 7 and 10 years of age. It is therefore possible that the authors of the studies with samples over 10 years old decided to apply the APBS’s adult version because of the absence of evidence on the psychometric properties of the children’s version for that age range. Our findings support the use of the APBS in teenage samples, given that the reliability estimate obtained in this sample was adequate (average alpha .884) according to the psychometric theory. However, it should be noted that the APBS total score reliability estimate might be increased with the participant’s age. In any case, future research should investigate the psychometric properties of children and adolescents’ APBS or even develop an adaptation of this scale for this age range.

Studies reporting and inducing the reliability of the APBS scores were compared for the sociodemographic characteristics of their samples to determine whether the conclusions of this RG meta-analysis can be generalized to other studies, and we found no statistically significant differences. If studies that reported and induced reliability used participant samples of similar composition and variability, the results of this RG meta-analysis can be generalized to any study that applies the APBS, regardless of whether they reported or induced their test score reliability.

Limitations

This RG meta-analysis has certain limitations that should be mentioned. First, language limitations could have affected the possibly eligible studies that applied the APBS. Second, the small number of studies that reported test–retest correlations could limit the generalizability of the results on the temporal stability of APBS scores. Third, due to the few studies that reported McDonald’s Omega and test–retest coefficients, it was not possible to carry out moderator analyses to search for potential variables related to heterogeneity shown by these reliability estimates.

Implications for Future Research and Clinical Practice

As reliability is not an inherent property of the test but depends on the composition and variability of the sample to which the test is applied, it is advisable to estimate and report reliability estimates with the data at hand, in line with the recommendations of international organizations, such as the American Psychological Association, the American Educational Research Association, and the National Research Council on Measurement in Education. This information would help researchers and professionals to determine the accuracy of the measurement and thus would allow them to extract conclusions based on this reliability.

In clinical practice, the APBS has shown acceptable average internal consistency reliability for clinical use. However, professional practitioners should consider that the future values of internal reliability estimates in other applications of the test might not be adequate for clinical practice, given that they might range between .746 and .957 (Nunnally & Bernstein, 1994). In addition, they should be aware that test–retest reliability showed that the construct measured by the APBS is not very stable and may change over time (e.g., with age, experience, learning, or the evaluation endeavor).

Declaration of Conflicting Interests

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

Ethical Statement

Ethical approval was not required due to the study being designed as a meta-analysis.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Informed Consent

There was no need for informed consent for the meta-analysis.

Research Involving Human Participants and/or Animals

This research has not done with human participants and/or animals but was designed as a meta-analysis.

ORCID iD

Laura Badenes-Ribera https://orcid.org/0000-0002-4706-690X

Data Availability

The data sets generated during and/or analyzed during the current study are available as Supplemental Material in journal's website.

Supplemental Material

Supplemental material for this article is available online.

References

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*(1), 3–25. https://doi.org/10.1037/amp0000389

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement*, *63*(1), 75–95.

Biagioli, V., Prandi, C., Giuliani, L., Nyatanga, B., & Fida, R. (2016). Prosocial behaviour in palliative nurses: Psychometric evaluation of the prosociality scale. *International Journal of Palliative Nursing*, *22*(6), 292–298. https://doi.org/110.12968/ijpn.2016.22.6.292

Blázquez-Rincón, D., Durán, J. I., & Botella, J. (2022). The fear of COVID-19 scale: A reliability generalization meta-analysis. *Assessment*, *29*(5), 940–948. https://doi.org/10.1177/1073191121994164

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335–340. https://doi.org/10.3102/10769986027004335

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*, 368–385. https://doi.org/10.1037/a0020142

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

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

Caprara, G. V., & Pastorelli, C. (1993). Early emotional instability, prosocial behaviour, and aggression: Some methodological aspects. *European Journal of Personality*, *7*(1), 19–36.

Caprara, G. V., Steca, P., Zelli, A., & Capanna, C. (2005). A new scale for measuring adult's prosocialness. *European Journal of Psychological Assessment*, *21*, 77–89. https://doi.org/10.1027/1015-5759.21.2.77

Carrizales, A., Perchec, C., & Lannegrand-Willems, L. (2019). Brief report: How many dimensions in the prosocial behavior scale? Psychometric investigation in French-speaking adolescents. *European Journal of Developmental Psychology*, *16*(3), 340–348. https://doi.org/10.1080/17405629.2017.1419952

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*(3), 290–304. https://doi.org/10.1080/00221300309601160

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*, *62*(5), 783–801. https://doi.org/10.1177/001316402236878

Egger, M., Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education*, *89*(1), 125–144. https://doi.org/10.1080/00220973.2019.1582470

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods*, *11*, 193–206. https://doi.org/10.1037/1082-989X.11.2.193

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*, 275–292. https://doi.org/10.1111/1468-2389.00156

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

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Harvard University Press.

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. https://doi.org/10.3102/1076998612466142

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. https://doi.org/110.1111/bmsp.12002

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

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). Russell Sage Foundation.

Regner, E., & Vignale, P. (2008). *Adaptación de la Escala de Conductas Prosociales de Caprara y Pastorelli* [Adaptation of the Caprara and Pastorelli Prosocial Behavior Scale] [Unpublished manuscript].

Revelle, W., & Condon, D. (2018). *Reliability from α to ω: A tutorial*. https://doi.org/10.31234/osf.io/2y3w9

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods*, *11*(3), 306–322. https://doi.org/10.1037/1082-989X.11.3.30

Rubio-Aparicio, M., Badenes-Ribera, L., Sánchez-Meca, J., Fabris, M. A., & Longobardi, C. (2020). A reliability generalization meta-analysis of self-report measures of muscle dysmorphia. *Clinical Psychology: Science and Practice*, *27*(1), 1–24. https://doi.org/10.1111/cpsp.12303

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

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

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibáñez, C., & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*, *12*(4), 516–536. https://doi.org/10.1002/jrsm.1487

Sawilowsky, S. S. (2000). Psychometrics versus datametrics: Comment on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Educational and Psychological Measurement*, *60*(2), 157–173. https://doi.org/10.1177/00131640021970439

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*(10), 1046–1055. https://doi.org/10.1016/S0895-4356(01)00377-8

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*(1), 6–20. https://doi.org/10.1177/0013164498058001002

Vacha-Haase, T., Henson, R. K., & Caruso, J. C. (2002). Reliability generalization: Moving toward improved understanding and use of score reliability. *Educational and Psychological Measurement*, *62*, 562–569. https://doi.org/10.1177/0013164402062004002

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*, 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metaphor package. *Journal of Statistical Software*, *36*(3), 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*, 360–374. https://doi.org/10.1037/met0000023

Ward, S. J., & King, L. A. (2018). Religion and moral self-image: The contributions of prosocial behavior, socially desirable responding, and personality. *Personality and Individual Differences*, *131*, 222–231. https://doi.org/10.1016/j.paid.2018.04.028

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*(4), 319–350. https://doi.org/10.1016/j.jrp.2004.03.001