Assessment 29.5 (2022): 940 – 948

DOI: https://doi.org/10.1177/1073191121994164

The Fear of Covid-19 Scale: A Reliability Generalization Meta-analysis

Desirée Blázquez

Universidad de Murcia, Murcia, Spain

Juan I. Durán

Universidad a Distancia de Madrid, Madrid, Spain

Juan Botella

Universidad Autónoma de Madrid, Madrid, Spain

Author's Note

This work was supported by the Ministerio de Economía, Industria y Competitividad of Spain [Project PSI2017-82490-P]. Thanks are due to James F. Juola, Professor Emeritus of Cognitive Psychology at the University of Kansas, who helped to polish the draft.

Any message regarding this article should be sent to the Division of Basic Psychology and Methodology, Department of Psychology, University of Murcia (Campus Universitario de Espinardo, 30100 Murcia, Spain).

Corresponding author: Desirée Blázquez

E-mail address: desireem.blazquez@um.es

Abstract

A reliability generalization meta-analysis was carried out to estimate the average reliability of the 7-item, 5-point Likert-type Fear of Covid-19 Scale (FCV-19S), one of the most widespread scales developed around the COVID-19 pandemic. Different reliability coefficients from Classical Test Theory and the Rasch Measurement Model were meta-analyzed, heterogeneity among the most reported reliability estimates was examined by searching for moderators, and a predictive model to estimate the expected reliability was proposed. At least one reliability estimate was available for a total of 44 independent samples out of 42 studies, being that Cronbach’s alpha was most frequently reported. The coefficients exhibited pooled estimates ranging from .85 to .90. The moderator analyses led to a predictive model in which the standard deviation of scores explained 36.7% of the total variability among alpha coefficients. The FCV-19S has been shown to be consistently reliable regardless of the moderator variables examined.

Keywords: FCV-19S, meta-analysis, reliability generalization, fear, Covid-19, classical test theory, Rasch measurement model.

The Fear of Covid-19 Scale: A Reliability Generalization Meta-analysis

One of the most widespread scales developed around the COVID-19 pandemic is the Fear of Covid-19 Scale (FCV-19S). Fear is a primitive and natural response that serves as a warning of imminent, real or perceived threats, and generally drives responses towards self-protection (Hooley et al., 2016). The original version of the FCV-19S was developed in Persian (Ahorsu et al., 2020). After debugging the primary bank of items, the final version consists of 7 Likert-type items with 5 categories ranging from "totally disagree" (1) to "totally agree" (5). Since its publication, its use has spread rapidly, being translated into multiple languages. Various studies have provided independent estimates of its psychometric properties obtained with samples of a generally reasonable size from diverse target populations.

The FCV-19S was designed to be a one-dimensional scale, and the results from the original and many subsequent studies support that claim. However, some controversy arises on this issue since other studies have suggested two factors (i.e., psychological and physiological symptoms) (Barrios et al., 2020; Huarcaya-Victoria et al., 2020; Masuyama et al., 2020; Reznik et al., 2020; Sugawara et al., 2020). As the evidence on its structure is not conclusive, and since most studies assume a single factor and report reliability coefficients for the full scale, we will treat it here as a one-dimensional scale.

Reliability generalization (RG) studies are a type of meta-analysis that quantitatively synthesize estimates of the reliability of the scores obtained with a given tool (Vacha-Haase, 1998). Reliability refers to the stability of the measurements of an unchanged trait or ability (Muñiz, 1992), and it depends on the composition and characteristics of the samples of participants, the contexts of the data collection, as well as the number of items and response categories of the test. As a consequence, the best way to guide expectations about the reliability of the scores is to quantitatively integrate several reliability estimates obtained from different administrations of the instrument. Meta-analysis is an optimal method to synthesize and examine how score reliability varies along different test applications (Henson and Thompson, 2002; Sánchez-Meca et al., 2013).

The goal of the present RG meta-analysis is to obtain an estimate of the average reliability of the FCV-19S by testing several characteristics of the studies as potential moderators that may explain part of the heterogeneity found among the collected reliability coefficients. The analysis of the moderator variables plays a somewhat different role from that of traditional meta-analyses. In general, significant moderators provide an opportunity to address and study new questions. They also allow us to make better decisions about interventions. On the contrary, the absence of significant moderators in RG studies implies that the combined reliability is uniform for the dimensions studied in the moderators. Therefore, multiple instruments are not required for reaching different levels of reliability as functions of the characteristics of participants or context. We also investigate the generalizability of our results by comparing the sample characteristics of the studies that made at least one reliability estimate available and those that did not.

Method

Search and Data Sources

Given the novelty of the FCV-19S, two different searches of relevant studies were made. In an initial search, the following data bases were consulted: APA (PsycInfo, PsycArticles and PsycBooks), Psychology and Behavioral Sciences Collection, PSICODOC, Academic Search Premier, MEDLINE, Google Scholar, PubMed, and Scopus. Because the FCV-19S original article was published online on May 27th of 2020, the search period of relevant studies covered that date until July 14th. In the search, the following keywords were set to be found anywhere in the documents: “The Fear of Covid-19 Scale” or “FCV-19S” or “FCV19S”. On July 17th, a second search was performed through Google Scholar with the aim of locating all the articles that cite any of studies found in the first search period. Those new reports that had not been located in the initial search, were checked to identify additional studies.

Inclusion and Exclusion Criteria

Three inclusion criteria were applied: (a) to be an empirical study where the FCV-19S, or a translation of this scale, was applied to a sample of more than one subject; (b) to report any reliability estimate computed with the data from the study-specific sample (focused on any target population); (c) to be written in English (at least, the abstract). The following exclusion criteria were applied: (a) to be a single subject clinical trial, or a case series study; (b) to have applied any other version of the FCV-19S that did not maintain the 7-item, 5-point Likert-type scale structure. The flowchart presented in Figure S1 of the supplemental material describes and adds details on the selection process of the studies. At the end of the process we had located 44 independent samples among 42 different studies in which any estimate of the reliability of the FCV-19S was reported in the own document, informed via email, or computed with the online dataset.

Data Extraction

Several moderator variables were coded concerning methodological aspects, sample characteristics, and COVID-19 related information. When data needed for coding the potential moderator variables was not reported, emails were sent to authors asking for that specific information.

Regarding the methodological aspects, the following variables were coded: (a) whether the study was published or not at the time of the second search; (b) whether the focus of the study was psychometric or applied; (c) language in which the FCV-19S was applied; (d) exact dates of the beginning and the end of data collection period; (e) factor structure of the FCV-19S based on the study-specific data; and (f) availability of the study's dataset.

With respect to the sample characteristics, we coded: (g) geographical location (country and continent); (h) target population (community, high school students, university students, clinical, and hospital staff); (i) sample size; (j) gender distribution of the sample (% male); (k) mean and standard deviation of the participants' ages (in years); and (l) mean and standard deviation of the FCV-19S total scores.

Along with all these characteristics, data related to the impact of covid-19 disease at the moment of data collection was also coded: (m) rate of deaths per confirmed cases in the country where data collection was conducted; (n) rate of recovered per confirmed cases; (o) rate of deaths per country population; (p) rate of recovered cases per country population; and (p) rate of confirmed cases per country population. Data for each population's census were extracted from the webpage of Country Meters (2020). Regarding covid-19 impact, data for the confirmed, deceases, and recovered cases for each study were extracted from the online GitHub hosting of the data for the COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University (2020; Dong et al., 2020). Given that for most of the studies data collection was conducted in more than one day, the exact date for coding the covid-19 impact was the intermediate day of the data collection period. For those studies that reported the relevant information for a sample composed of participants from two different countries, the moderator variables related to the impact of COVID-19 were computed as the weighted average of each country's sample size.

The reliability of the coding process of the study characteristics was checked by doubly coding 23 (50%) of studies that had applied the FCV-19S by the first and second authors, both psychologists with experience in meta-analysis. In general, the inter-coder agreement was satisfactory, with kappa coefficients ranging between .78 and 1.00 for the qualitative characteristics, and intraclass correlations ranging between .75 and 1.00 for the continuous variables. Inconsistencies between both coders were solved by consensus.

Although only those studies for which at least one reliability estimate was available were included in the main analysis, all potential moderators were also extracted from those studies for which reliability remained unknown in order to examine the generalizability of our results.

Statistical Analyses

Separate meta-analyses were carried out for those types of reliability coefficients reported in at least five independent samples (Dimitrov, 2002; Sawilowsky, 2000). Formulas and references for the meta-analyzed reliability indices can be found in Table S1 of the supplemental material. Regardless of the kind of reliability involved, all coefficients (ρ_{XX}) were transformed to normalize their distributions and stabilize their variances following the recommendations by Sánchez-Meca et al. (2013). Due to the natural asymmetry of their distributions, reliability coefficients should be transformed in order to meet the assumptions of the analyses used to meta-analyze them and search for moderators. For those indices that are based on Pearson correlation coefficient and range between -1 and 1, like split-half reliability coefficients, the most suitable transformation is Fisher's Z: $T_F = 0.5[Ln(1 + \rho_{XX}) - Ln(1 - \rho_{XX})]$. Whereas, for those reliability coefficients that range between 0 and 1 (Cronbach's alpha, McDonald's Omega, Guttman's lambda 2 and 6, composite reliability, greatest lower bound reliability, and person separation reliability), a theoretically more appropriate transformation is the one proposed by Bonett's (2002): $T_B = Ln(1 - \rho_{XX})$, where Ln is the natural logarithm.

As random-effects models were assumed, the reliability coefficients were weighted by the inverse variance method, where the variance is the sum of the within-study and the between-studies variances. Between-study variance, τ^2, was estimated using the Paule and Mandel estimator (Boedeker and Henson, 2020). The 95% confidence interval around each overall reliability estimate was computed with the improved method proposed by Hartung (1999). Heterogeneity among estimates of the same reliability coefficient was assessed with the Q test and the I^2 index, assuming a level of significance of .05 and I^2 values of approximately 25%, 50%, and 75% as low, moderate, and large heterogeneity, respectively (Higgins et al., 2003).

For meta-analyses with at least 30 reliability estimates, moderator analyses were conducted through meta-regression analyses for continuous variables and weighted analysis of variance (ANOVA) for qualitative variables. Mixed-effects models were assumed, using the improved F statistic proposed by Knapp and Hartung (2003) to test the statistical significance of moderator variables, given that it offers a better control of the Type I error rate than the standard Q_B and Q_R statistics (Hedges and Olkin, 1985). The Q_E and Q_W statistics allow testing the model misspecification for meta-regression and ANOVA, respectively. The proportion of variance accounted for by the moderator variables was estimated with R^2, as it considers the total and residual between-study variances (López-López et al., 2014). All statistical analyses were carried out with the metafor package in R (Viechtbauer, 2010), whereas for the calculation of the reliability coefficients, the packages eRm, dplyr, lavaan, psy, psych, Rcsdp, and semTools were used.

Finally, additional meta-analyses were carried out to compare the sample characteristics of those studies for which at least one reliability estimate was known and those for which it was not. Specifically, the particular characteristic was used as the effect size measure in a new mixed-effects meta-analysis, where reliability induction (a categorical variable indicating whether reliability was known, “1”, or induced, “0”) acts as the single moderator variable.

Complete data and R code files used for the analyses reported below, alongside the supplemental material, are available at Mendeley (Blázquez-Rincón et al., 2020).

Results

Descriptive Characteristics of the Studies

The references for the included studies are specified in the supplemental material. The number of independent estimates, its percentage in relation to the 44 samples included in the study, and the cumulated sample size are in Table 1. The total sample was N = 46,223. The distribution of sample sizes was extremely right-skewed and leptokurtic, with a mean of 1050.5 participants per sample (median = 611, SD = 12,633.34, skewness = 14.82, kurtosis = 229.53). The average male percentage was 37% (SD = 18), and the average of the participants' mean age and its standard deviation equaled 32.6 and 10.4, respectively. The average of mean and standard deviation of the FCV-19S total scores were 18.36 and 5.9, respectively.

Mean Reliability and Heterogeneity

Separate meta-analyses were conducted for each type of reliability coefficient. Although the statistical analyses were performed using different transformations for raw reliability estimates, all tables and figures show the pooled means and their respective confidence limits once back transformed to the corresponding coefficient metric for the purpose of facilitating interpretation.

The pooled reliability for each type of coefficient and the 95% confidence interval, alongside the *Q* test and the I^2 index can be found in Table 1. The results for the most reported reliability estimate, Cronbach's alpha, yielded a mean coefficient of .865, ranging from .79 to .93. The combined estimates for the different reliability coefficients were very close (from .850 to .901). Evidence of heterogeneity was found for all the reliability coefficients, with all *Q* statistics being significant ($p < .001$). I^2 indexes were of large magnitude in all cases, ranging from 89.74 to 97.87%. Table 1 also includes the lower limit of the credibility interval. The values reflect the estimated thresholds that would be reached by each coefficient in about 90% of future applications of the scale.

Analyses of Moderator Variables

Moderator analyses were only carried out to explain the large variability exhibited by the alpha coefficients, since is the only one for which we have obtained at least 30 reliability estimates. Table 2 shows the results of the simple meta-regressions applied on alpha coefficients for the FCV-19S. Among the different quantitative moderators analyzed, the standard deviation of test scores was the only one which exhibited a positive, statistically significant relationship with the coefficient alpha, accounting for 36.7% of the variance.

With regard to qualitative moderators, Table 3 presents the results of the ANOVAs. Due to the large number of translations of the original FCV-19S (in Persian) to at least eighteen different languages, those languages for which at least five reliability estimates were unavailable were combined into a single category. Only the target population of the studies showed a statistically significant relationship with the alpha coefficients ($p = 0.037$). This moderator accounts for 17% of the variance, with a larger average alpha coefficient for studies conducted with clinical samples related with mental disorders.

Explanatory Models

Although two moderators showed a statistically significant association with the alpha coefficients, all misspecification tests suggested residual heterogeneity among the reliability estimates (see Tables 2 and 3). Thus, we have used multiple meta-regression to identify the most relevant study characteristics to explain the variability.

Due to theoretical reasons, there is an expected relationship between the reliability of a scale and the variability among its total scores. Therefore, when carrying out analysis for multiple moderator variables in RG meta-analyses, the variability of test scores must be included as a predictor in first place (Botella et al., 2010). The explanatory model examined in the first stage included five predictors: the standard deviation of the FCV-19S total scores and the target population of the study, coded as a set of four dummy variables for general, university students, high school students, and hospital staff. Before fitting our explanatory model, we concluded that there was no relationship among the predictors after carrying out a Kruskal Wallis non-parametric test. Due to missing data in some variables, reliability estimates for only 41 independent samples were included in the model. The initial model showed a statistically significant relationship with the alpha coefficients ($F(6, 35) = 5.92$, $p < .001$), accounting for 42.83% of the variance. Of both predictors included in the model, only the standard deviation of the total scores showed a statistically significant relationship with the alpha coefficients ($F(1, 35) = 14.58$, $p < .001$) once the influence of the other variable was controlled. The relationship between alpha coefficients and the target population did not yield statistically significant results ($F(4, 35) = 1.92$, $p = 0.129$). As related to the contribution in terms of proportion of variance increased by each predictor to the multiple meta-regression model, the inclusion of the target population led to an increase of 6.17% of accounted variance, once the remaining predictor had already been included in the model.

After removing the target population as a predictor variable, the final explanatory model only includes the standard deviation of total scores as a predictor, accounting for 36.7% of the variability showed among the alpha coefficients.

Comparing Studies that Induce and Report/Inform Reliability

Despite the recommendations of the American Psychological Association (Appelbaum et al., 2018), not all the primary studies reported a reliability coefficient. Of the 44 studies that had applied the FCV-19S at the time of the second search, 8 induced reliability, implying a 18.18% reliability induction. Of the 8 studies that induced reliability, 4 (50%) omitted any reference to the FCV-19S reliability (induction by omission), whereas the remaining 4 studies (50%) induced reliability by reporting a reliability estimate computed in a different sample (induction by report). Reliability estimates for 6 studies out of those that had induced reliability were available for the main analysis, but data for the other 2 remain unknown. Overall, 7 independent samples were extracted from the 2 studies that did not report/inform any reliability coefficient.

In order to examine to what extent our results can be generalized to the whole sample of studies that applied the FCV-19S, we compared the sample characteristics of the studies that made a reliability estimate computed with their own data available in their article or via e-mail, and those studies that did not. Results of these comparisons can be found in Table 4. The mean age of the studies that did not report the reliability with their own data was lower than that of studies that reported reliability with the data at hand ($p < .005$), whereas the mean total scores in FCV-19S was higher in those studies that reported the reliability with their own data ($p < .004$). There was a marginally significant difference between the variances of participant’s age ($p =$

.0863): those studies that reported the reliability with their own data showed a higher age variance than those that did not. However, there were no differences between the variances of FCV-19S scores.

Discussion

The present RG meta-analysis on the Fear of COVID-19 Scale examines how its reliability estimates vary through different administrations, and guides reliability expectations for future applications. The most commonly reported reliability coefficient has been Cronbach's alpha. Still, some studies also reported other reliability coefficients or posted their data online, so we were able to obtain an average reliability of the FCV-19S based on those other coefficients for which at least five estimates were available. The different reliability coefficients exhibited pooled estimates that were quite homogeneous, ranging from .850 to .901. Most of the guidelines for interpreting the adequacy of reliability estimates are based on the alpha coefficient, so we must be cautious in applying them to the rest of the meta-analyzed coefficients. Tentatively, on average, the reliability of FCV-19S was clearly above the cutoff of .70, considered as the minimum recommended when a test is administered with exploratory research purposes, and over the threshold of 0.80, recommended for general research purposes (Nunnally and Bernstein, 1994). Nevertheless, only the greatest lower bound coefficient reached the stricter criterion of .90 established for important clinical decisions derived from the scaled scores (Cicchetti, 1994).

The several reliability coefficients reported in the present meta-analysis represent different ways of measuring reliability within Classical Test Theory (CTT). Although the reliability concept is fundamentally different between Item Response Theory and CTT, the concept of Person Separation Reliability, similar to previous reliability coefficients, is understood as the proportion of person variance that is not due to error. As all of these coefficients vary in their calculation, they also vary in the extent to which they are affected by different measurement errors. Regardless of the differences in the number of estimates available for each reliability coefficient, the coefficient that yielded the lowest pooled estimate was the split-half reliability, whereas the greatest lower bound coefficient reached the highest average. These results could be due to the fact that split-half reliability is based on the correlations between split halves that are assumed to be parallel when, actually, the FCV-19S is made up of 7 items which makes it impossible to get two truly parallel halves. Regarding the greatest lower bound, this coefficient can be interpreted as the lowest possible value that a scale's reliability can reach, similarly to Cronbach's alpha and McDonald's omega, but it has been shown to produce better results over alpha and omega even in non-normal conditions (Trizano-Hermosilla and Alvarado, 2016).

Concerning the analyses for the moderators, only one variable explained a significant part of the heterogeneity shown by the alpha estimates: the variability of the FCV-19S scores. Those samples where the standard deviation of the total scores was higher, also showed a higher alpha coefficient. The variability among the scores of a test is expected to be related to the test reliability. The median percentage of explained variance by the standard deviation of scores in those RG meta-analyses collected by López-Ibáñez et al. (2020) was 41.97 (*SD* = 29.29), whereas the percentage found for the FCV-19S was 36.1%. According to Botella et al. (2010), the relationship between the variability among the observed scores and the corresponding reliability coefficient is complex, since it may be direct (e.g., when the variability of the variances is due to differential sampling schemes) or inverse (when it is due to changes in the measurement errors).

Finally, one aspect of RG meta-analyses that distinguishes it from other kinds of meta-analyses, is the potential impact of what has been called reporting bias. Strictly speaking, there is no publication bias in the field of reliability generalization, as reliability estimates are not reported in terms of their statistical significance. Nevertheless, as stated above, there are cutoff points that serve as guides for interpreting most of the reliability coefficients. This may lead some researchers to choose whether to report a reliability estimate concerning the scales applied in their studies. In the present meta-analysis, we have not found differences between the studies that report and those that do not report it in the characteristics found to be significant moderators.

Limitations and Future Research

The main limitation of our study comes from the novelty of the scale. Despite the fact that the original FCV-19S article was published on May 27th (2020), a huge number of studies have applied it to this moment. However, other studies in which the FCV-19S is planned to be applied in future stages of the COVID-19 pandemic were not part of the present meta-analysis. Still, a reliability generalization meta-analysis on the FCV-19S is in order due to the need to develop and study the psychometric properties of new assessment tools specifically designed to identify and study mental health issues related to COVID-19.

Another limitation is the scarce number of studies that reported results concerning exploratory factor analyses (EFA) of the FCV-19S. Most of the studies that have been carried out concerning confirmatory factor analyses have fit the one-factor structure. However, although the fit indices are acceptable, it is recommended to search for other underlying structures that may improve the data fit. We encourage researchers to perform EFA of different factor structures when applying the FCV-19S, given that reliability may be affected by the underlying structure of the scale.

Another controversial issue is whether there are differences between men and women in the properties of the FCV-19S. Several studies have found that women get higher total scores compared to men, whereas other studies have not. With the data at hand, we can state that there is no evidence of differences between men and women regarding either the internal consistency of the FCV-19S, or the total scores (r = 16.8, p = .283). However, further research should be carried out to get a more comprehensive overview of any gender differences in the psychometric properties of the FCV-19S.

References

Ahorsu, D. K., Lin, C. Y., Imani, V., Saffari, M., Griffiths, M. D., & Pakpour, A. H. (2020). The fear of COVID-19 scale: development and initial validation. *International Journal of Mental Health and Addiction*. Doi: 10.1007/s11469-020-00270-8.

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. Doi: 10.1037/amp0000389.

Barrios, I., Ríos-González, C., O'Higgins, M., González, I., García, O., Ruiz-Díaz, N., Castaldelli-Maia, J. M., Ventriglio, A., & Torales, J. (2020). *Psychometric properties of the Spanish version of the Fear of COVID-19 Scale (FCV-19S)*. Unpublished manuscript. Doi: 10.21203/rs.3.rs-33345/v1.

Blázquez-Rincón, D., Durán, J. I., & Botella, J. (2020). *The Fear of COVID-19 Scale: A Reliability Generalization Meta-analysis*, Mendeley Data, V1. Available at http://dx.doi.org/10.17632/dtwdf566xm.1

Boedeker, P., & Henson, R. K. (2020). Evaluation of heterogeneity and heterogeneity interval estimators in random-effects meta-analysis of the standardized mean difference in education and psychology. *Psychological Methods*, *25*(3), 346–364. Doi: 10.1037/met0000241.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of educational and behavioral statistics*, *27*(4), 335-340. Doi: 10.3102/10769986027004335.

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. Doi: 10.1037/a0019626.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessments instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. Doi: 10.1037/1040-3590.6.4.284.

Country Meters (2020). *Population of the World and Countries.* Retrieved July 20th, 2020 from https://countrymeters.info/en

Center for Systems Science and Engineering (2020). *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. Johns Hopkins University. Retrieved July 20th, 2020 from https://github.com/CSSEGISandData/COVID-19

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

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533-534. Doi: 10.1016/S1473-3099(20)30120-1.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *41*(8), 901-916. Doi: 10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods in meta-analysis* (1st ed.). Academic Press.

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*(2), 113-127. Doi: 10.1080/07481756.2002.12069054.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*(7414), 557–560. Doi: 10.1136/bmj.327.7414.557.

Hooley, J. M., Butcher, J. N., Matthew, K. N., & Mineka, S. (2016). *Abnormal psychology* (17th ed.). Pearson.

Huarcaya-Victoria, J., Villarreal-Zegarra, D., Podestà, A., & Luna-Cuadros, M. A. (2020). Psychometric properties of a Spanish version of the Fear of COVID-19 Scale in general population of Lima, Peru. *International Journal of Mental Health and Addiction*. Doi: 10.1007/s11469-020-00354-5.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*(17), 2693-2710. Doi: 10.1002/sim.1482.

López-Ibáñez, C., López-Nicolás, R., Blázquez-Rincón, D., & Sánchez-Meca, J. (2020). *Reliability generalization meta-analysis: A comparison of statistical analytic strategies*. Unpublished manuscript.

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*(1), 30-48. Doi: 10.1111/bmsp.12002.

Masuyama, A., Shinkawa, H., & Kubo, T. (2020). Validation and Psychometric Properties of the Japanese Version of the Fear of COVID-19 Scale Among Adolescents. *International Journal of Mental Health and Addiction*. Doi: 10.1007/s11469-020-00368-z.

Muñiz, J. (1992). *Teoria clásica de los tests* (1st ed.). Pirámide.

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

Reznik, A., Gritsenko, V., Konstantinov, V., Khamenka, N., & Isralowitz, R. (2020). COVID-19 fear in Eastern Europe: Validation of the Fear of COVID-19 Scale. *International Journal of Mental Health and Addiction*. Doi: 10.1007/s11469-020-00283-3.

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

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. Doi: 10.1177/00131640021970439.

Sugawara, D., Masuyama, A., & Kubo, T. (2020). *Socioeconomic Impacts of the COVID-19 Lockdown on the Mental Health and Life Satisfaction of the Japanese Population*. Unpublished manuscript. Doi: 10.31234/osf.io/sndpm.

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. *Frontiers in psychology*, *7*, 769. Doi: 10.3389/fpsyg.2016.00769.

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

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3). Doi: 10.18637/jss.v036.i03.

Table 1

Average reliability coefficients, 95% confidence intervals, lower limit of the 80% credibility interval, and heterogeneity statistics for the Fear of Covid-19 Scale (FCV-19S).

Reliability estimate	***k* (%)**	**N**	ES_+	**95% CI**	**LL 80% CRI**	***Q***	I^2
Classical Test Theory							
Cronbach's alpha	43 (97.73)	45,932	.865	[.855, .874]	.820	749.940***	95.71
McDonald's Omega	13 (29.55)	10,335	.857	[.820, .886]	.769	492.790***	97.87
Guttman's lambda 2	7 (15.91)	5,538	.853	[.820, .880]	.808	96.788***	93.29
Guttman's lambda 6	8 (18.18)	6,842	.853	[.823, .879]	.806	133.442***	94.28
Composite reliability	14 (31.82)	17,319	.854	[.830, .874]	.798	313.774***	96.33
Greatest lower bound	8 (18.18)	5,766	.901	[.879, .919]	.867	115.530***	93.74
Split-half reliability	8 (18.18)	6,206	.850	[.821, .875]	.812	73.172***	89.74
Rasch Measurement Theory							
Person Separation reliability	11 (25.00)	15,128	.854	[.821, .881]	.787	218.562***	97.34

Note. k = number of independent samples were the estimate of that particular reliability coefficient was available; N = total sample size; ES_+ = pooled reliability estimate; CI = confidence interval; LL = lower limit; CRI = credibility interval; Q = Cochran's heterogeneity Q statistic with k-1 degrees of freedom; I^2 = heterogeneity index.

*** $p < .001$

Table 2

Results of the simple meta-regressions applied on coefficient alphas for the FCV-19S, taking continuous moderator variables as predictors.

Predictor variable	***k* (%)**	**N**	b_j	***F***	***p***	Q_E	R^2
Sample size	43 (100.00)	45,932	.000	0.840	.365	679.837***	.000
Gender (% male)	43 (100.00)	45,932	.003	2.134	.152	694.097***	.030
Age (years)							
Mean	42 (97.67)	45,431	.008	2.321	.136	741.415***	.035
Standard deviation	41 (95.35)	45,203	.001	0.028	.869	661.174***	.000
FCV-19S total score							
Mean	43 (100.00)	45,932	-.007	0.285	.596	716.858***	.000
Standard deviation	41 (95.35)	43,979	.173	20.406	<.001	664.306***	.367
COVID-19 information							
Per infected cases							
Deaths	38 (88.37)	41,609	1.518	1.037	.315	721.116***	.011
Recovered	38 (88.37)	41,609	0.117	0.727	.399	607.120***	.000

Per population in 2020							
Deaths	38 (88.37)	41,609	635.792	.532	.471	720.096***	.000
Recovered	38 (88.37)	41,609	-69.885	.776	.384	701.028***	.000
Infected	38 (88.37)	41,609	0.112	.000	.998	698.136***	.000

Note. k = number of independent samples with respect to the 43 samples for which alpha coefficients were available; N = total sample size; bj = regression coefficient of each predictor; F = Knapp-Hartung's statistic for testing the significance of the predictor, with 1 degrees of freedom 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.

*** $p < .001$

Table 3

Results of the weighted analyses of variance (ANOVAs) applied on coefficient alphas for the FCV-19S, taking qualitative moderator variables as independent variables.

Predictor variable	***k* (%)**	**N**	α_+	**95% CI**	**ANOVA results**
Published					
Yes	29 (67.44)	34,321	.864	[.851, .876]	$F(1, 41) = 0.0734$, $p = .7877$. $R^2 = .000$
No	14 (32.56)	11,611	.867	[.848, .883]	$Q_W(41) = 746.4383$, $p < .001$
Study focus					
Psychometric	19 (44.19)	27,098	.871	[.856, .884]	$F(1, 41) = 1.2973$, $p = .2613$. $R^2 = .008$
Applied	24 (55.81)	18,833	.860	[.845, .873]	$Q_W(41) = 736.5364$, $p < .001$
Language					
English	7 (16.28)	3,686	.871	[.844, .893]	$F(3, 39) = 0.0964$, $p = .9616$
Turkish	6 (13.95)	6,082	.865	[.835, .889]	$R^2 = .000$
Spanish	5 (11.63)	6,756	.862	[.828, .889]	$Q_W(39) = 739.2568$, $p < .001$
Other	25 (58.14)	29,407	.864	[.850, .877]	
Factor structure					

One-factor	17 (39.53)	23,243	.877	[.863, .891]	*F*(4, 38) = 1.5546, *p* = .2062.
Two-factors	2 (4.65)	1,927	.835	[.773, .881]	R^2 = .057
Bifactor	3 (6.98)	2,014	.852	[.807, .887]	Q_W(38) = 685.8188, *p* < .001
Neither one nor two-factors	1 (2.33)	370	.842	[.746, .902]	
Not stated	20 (46.51)	18,378	.859	[.843, .873]	
Data availability					
Yes	7 (16.28)	5,538	.844	[.815, .869]	*F*(1, 41) = 3.1647, *p* = .0827. R^2 = .056
No	36 (83.72)	40,394	.868	[.858, .878]	Q_W(41) = 707.5913, *p* < .001

Predictor variable	***k* (%)**	**N**	α_+	**95% CI**	**ANOVA results**
Continent					
America	7 (16.28)	8,838	.872	[.847, .893]	*F*(2, 40) = 1.4365, *p* = .2498
Asia	28 (65.12)	30,012	.868	[.855, .879]	R^2 = .023
Europe	8 (18.60)	7,082	.847	[.819, .871]	Q_W(40) = 681.5982, *p* < .001
Target population					

General	31 (72.09)	32,789	.862	[.850, .872]	$F(4, 38) = 2.8523$, $p = .0367$
University students	7 (16.28)	7,784	.866	[.842, .887]	$R^2 = .170$
Hospital staff	3 (6.98)	4,337	.876	[.840, .903]	$Q_W(38) = 539.9001$, $p < .001$
High School students	1 (2.33)	622	.822	[.725, .885]	
Clinical	1 (2.33)	400	.930	[.891, .955]	

Note. k = number of independent samples with respect to the 43 samples for which alpha coefficients were available; N = total sample size; α_+= pooled Cronbach's alpha estimate; 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 weighted analyses of variance (ANOVAs) applied on different sample characteristics of the studies that reported and induced test score reliability.

Sample characteristic	Average	95% CI	ANOVA results
Age			
Mean:			
Induced	24.13	[18.73, 29.53]	$F(1, 47) = 8.7401$, $p = .0049$. $R^2 = .139$
Reported	32.70	[30.50, 34.91]	$Q_W(47) = 111{,}956.5$, $p < .0001$
Variance:			
Induced	57.06	[0.00, 133.35]	$F(1, 47) = 3.0691$, $p = .0863$. $R^2 = .044$
Reported	128.86	[97.58, 160.14]	$Q_W(47) = 21{,}202.39$, $p < .0001$
Total score			
Mean:			
Induced	21.98	[19.82, 24.14]	$F(1, 47) = 9.7200$, $p = .0031$. $R^2 = .155$
Reported	18.36	[17.48, 19.24]	$Q_W(47) = 10{,}658.34$, $p < .0001$

Variance:			
Induced	29.88	[22.89, 36.87]	$F(1, 47) = 1.7492$, $p = .1924$. $R^2 = .018$
Reported	34.82	[32.05, 37.60]	$Q_W(47) = 1{,}085.51$, $p < .0001$

Note. 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.