



# Internal consistency and test-retest reliability: A reliability generalization meta-analysis of the Childhood Trauma Questionnaire – Short Form (CTQ-SF)

Laura Badenes-Ribera<sup>a</sup>, Sylvia Georgieva<sup>a,\*</sup>, Jose M. Tomás<sup>a</sup>, José Javier Navarro-Pérez<sup>b</sup>

<sup>a</sup> Faculty of Psychology, University of Valencia, Av. Blasco Ibáñez, 21, Valencia 46010, Spain

<sup>b</sup> Faculty of Social Work, University of Valencia, Av. Dels Tarongers, 4b, Valencia 46022, Spain

## ARTICLE INFO

### Keywords:

CTQ-SF

Childhood maltreatment

Reliability

Generalization

Meta-analysis

## ABSTRACT

**Background:** The Childhood Trauma Questionnaire- Short Form (CTQ-SF) is among the most extensively studied and widely utilized instruments for evaluating childhood maltreatment.

**Objective:** The purpose of this study was to conduct a reliability generalization meta-analysis to estimate the average reliability of the CTQ-SF scores and its factors and search for study characteristics that can explain the variability in those coefficients.

**Methods:** A total of 39 independent samples provided 243 reliability estimates (Cronbach's alpha, McDonald's omega, and/or test-retest reliability coefficients) with the data at hand for the scores on the CTQ-SF and its five subscales for this meta-analysis.

**Results:** Random and mixed-effects models were employed for analyzing the data. The average Cronbach's alpha coefficient for the CTQ-SF total score was 0.891 (95 % CI: 0.868, 0.910). For the subscales, the average Cronbach's alpha coefficient ranged from 0.656 (Physical Neglect) to 0.916 (Sexual Abuse). The average McDonald's Omega coefficient for the CTQ-SF total score was 0.800 (95 % CI: 0.800 0.800). For the subscales, the average McDonald's Omega ranged from 0.740 (Physical Neglect) to 0.900 (Sexual Abuse). The average test-retest reliability for CTQ-SF total score was 0.788 (95 % CI: 0.635, 0.872), with the subscales ranging from 0.668 (Physical Neglect) to 0.709 (Physical Abuse). Moderator analyses revealed that some factors can affect reliability estimate.

**Conclusions:** Although CTQ-SF and its five subscales have shown adequate reliability, it may vary as a function of the variability of scores, geographical location, financial source, and the affiliation of the main researcher.

## 1. Introduction

Child maltreatment can be defined as a range of harmful behaviors and experiences that are inflicted upon minors. These experiences include but are not limited to, maltreatment, neglect, physical or emotional mistreatment, sexual abuse, negligence, exposure to violent scenes, and/or commercial or other exploitative practices that can result in actual or potential harm to the child or

\* Corresponding author.

E-mail address: [sylvia.georgieva@uv.es](mailto:sylvia.georgieva@uv.es) (S. Georgieva).

<https://doi.org/10.1016/j.chiabu.2024.106941>

Received 7 September 2023; Received in revised form 28 June 2024; Accepted 2 July 2024

Available online 17 July 2024

0145-2134/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

adolescent's life, health, development, or dignity (WHO, 2020). Child maltreatment is a highly prevalent, complex, and multidimensional construct with diverse risk factors, posing challenges to its study and measurement (Georgieva et al., 2021; WHO, 2020).

In a systematic review by Saini et al. (2019), evaluating the psychometric properties of 52 instruments, the Childhood Trauma Questionnaire (CTQ; Bernstein & Fink, 1998), Childhood Trauma Questionnaire Short Form (CTQ-SF; Bernstein et al., 2003), and Maltreatment and Abuse Chronology of Exposure (MACE; Teicher & Parigger, 2015) emerged as the tools with the strongest psychometric evidence. In this line, a subsequent review evaluating the five most used instruments from 2010 to 2020 identified MACE, Psychosocial Screening Tool (PAT), and CTQ-SF as having the strongest scientific evidence of good psychometric properties (Georgieva et al., 2022).

The CTQ-SF has been examined in both recent systematic reviews. This instrument is among the most extensively employed tools for evaluating childhood maltreatment. Indeed, it has been recognized as a leading measure for assessing this phenomenon (Baker & Maiorino, 2010; Schmidt et al., 2020). Additionally, the short version is the recommended version by Bernstein et al. (2003). The CTQ-SF is the shortened form of the original CTQ. It is a retrospective instrument designed to assess five types of childhood abuse and maltreatment: sexual abuse, physical abuse, emotional abuse, physical neglect, and emotional neglect. The original instrument is comprised of 70 items distributed across these five subscales. The CTQ-SF is shortened to 28 items, 5 items per subscale, and 3 items assessing minimization and denial. Each item is rated on a 5-point Likert scale, ranging from 1 (never true) to 5 (very often true), thus, higher scores indicate severe maltreatment. To develop a short version of the CTQ, Bernstein et al. (2003) performed both exploratory and confirmatory factor analyses. The main purpose of the CTQ-SF is to be a screening measure for maltreatment histories in both clinical and non-referred groups from 12 years onwards (Bernstein et al., 2003). It is a self-administered instrument, thus, the child or adult answers for himself or herself about his/her own experiences. This scale has extensively been used in research settings in all age ranges but especially in adult samples, however, the most common way to administer it in research is using the 25 items without including the 3 minimization and denial items (Georgieva et al., 2021).

The CTQ-SF has undergone extensive validations across a variety of samples, including adolescents, substance users, psychiatric inpatients (Bernstein et al., 2003), street youth (Forde et al., 2012), prisoners (Dovran et al., 2013), juvenile delinquents (Kongerslev et al., 2019), and pregnant women (Schmidt et al., 2020). Moreover, the scale has been tested for measurement invariance across four groups and for criterion validity, with favorable psychometric results (Bernstein et al., 2003). Virtually all studies conclude that the CTQ-SF is a valid and reliable instrument for use.

However, despite its solid psychometric properties, a recently conducted systematic review revealed that some aspects of the instrument require further examination and clarification (Georgieva et al., 2021). For instance, it has been shown that internal consistency in some factors is sub-optimal according to the Measurement Instruments (COSMIN) checklist for Patient-Reported Outcome

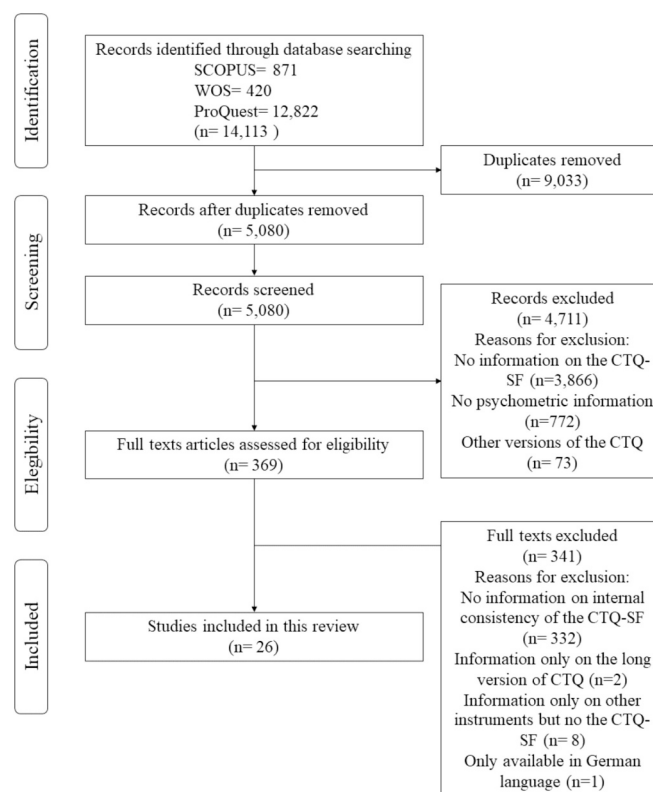


Fig. 1.. Flowchart of the studies selection process.

**Table 1**  
Characteristics of the studies (N = 25).

|    | Authors               | Year | Affiliation*   | Financed | Study language | Location    | Continent     | Research design | Sampling method | Sample size | Population target                                               | Mean Age (SD) | % male | % Caucasians | Language adaptation |
|----|-----------------------|------|----------------|----------|----------------|-------------|---------------|-----------------|-----------------|-------------|-----------------------------------------------------------------|---------------|--------|--------------|---------------------|
| 1  | Bernstein et al.      | 2003 | Psychologist   | No       | English        | USA         | North America | Cross-sectional | Not mentioned   | 396         | Adolescent psychiatric inpatients                               | 14.90 (1.40)  | 43.0   | 67.9         | Original            |
|    |                       |      |                |          |                |             |               |                 |                 | 378         | Adult substance abusers                                         | 40.20 (8.80)  | 85.6   | 13.4         |                     |
|    |                       |      |                |          |                |             |               |                 |                 | 579         | Community members                                               | 34.90 (NR)    | 28.0   | 67.0         |                     |
|    |                       |      |                |          |                |             |               |                 |                 | 625         | Adult substance abusers                                         | 34.00 (NR)    | 64.0   | 11.0         |                     |
| 2  | Gerdner & Allgulander | 2009 | Social worker  | No       | English        | Sweden      | Europe        | Cross-sectional | Not mentioned   | 420         | Clinical mental health population                               | 45.00 (7.40)  | 50.0   | NR           | Original            |
|    |                       |      |                |          |                |             |               |                 |                 | 239         | Nonclinical population                                          | 27.00 (7.40)  | 34.0   | NR           |                     |
| 3  | Thombs et al.         | 2009 | Psychiatrist   | Yes      | English        | Netherlands | Europe        | Longitudinal    | Not mentioned   | 488         | Clinical and non-clinical participants                          | 33.80 (12.00) | 43.2   | NR           | Dutch               |
| 4  | Bogaerts et al.       | 2011 | Psychologist   | No       | English        | Netherlands | Europe        | Cross-sectional | Stratified      | 123         | Female sex workers                                              | 33.2 (9.46)   | 0      | NR           | Dutch               |
| 5  | Kim et al.            | 2011 | Psychiatrist   | Yes      | English        | Korea       | Asia          | Cross-sectional | Convenience     | 163         | Psychiatric outpatients                                         | 38.30 (11.80) | 39.3   | NR           | Korean              |
| 6  | Forde et al.          | 2012 | Criminologist  | Yes      | English        | USA         | North America | Cross-sectional | Incidental      | 240         | Street youth male                                               | 20.00 (2.60)  | 100.0  | 83.0         | Original            |
|    |                       |      |                |          |                |             |               |                 |                 | 160         | Street youth female                                             | 20.00 (2.60)  | 0      | 83.0         |                     |
| 7  | Hernandez et al.      | 2013 | Psychiatrist   | Yes      | English        | Spain       | Europe        | Cross-sectional | Convenience     | 185         | Clinical mental health population                               | 41.64 (11.83) | 0      | NR           | Spanish             |
|    |                       |      |                |          |                |             |               |                 |                 | 72          | Foster care adolescents                                         | 19.10 (3.50)  | 44.0   | NR           |                     |
|    |                       |      |                |          |                |             |               |                 |                 | 126         | Mental health and addition patients                             | 30.90 (11.60) | 60.0   | NR           |                     |
|    |                       |      |                |          |                |             |               |                 |                 | 210         | Psychiatric patients                                            | 25.50 (10.70) | 30.0   | NR           |                     |
| 8  | Dovran et al.         | 2013 | Psychologist   | Yes      | English        | Norway      | Europe        | Cross-sectional | Not mentioned   | 109         | Prisoners                                                       | 32.20 (9.90)  | 100.0  | NR           | Norwegian           |
|    |                       |      |                |          |                |             |               |                 |                 |             |                                                                 |               |        |              |                     |
| 9  | Karos et al.          | 2014 | Psychologist   | No       | English        | Germany     | Europe        | Cross-sectional | Not mentioned   | 661         | Psychology students and mental health patients                  | 35.08 (11.75) | NR     | NR           | German              |
| 10 | Spinhoven et al.      | 2014 | Psychologist   | Yes      | English        | Netherlands | Europe        | Longitudinal    | Not mentioned   | 2308        | Controls, outpatients and inpatients for depression and anxiety | 46.00 (13.10) | 33.6   | NR           | Dutch               |
| 11 | Dudeck et al.         | 2015 | Psychiatrist   | No       | English        | Germany     | Europe        | Cross-sectional | Not mentioned   | 1524        | Adult Psychiatric patients                                      | 37.70 (12.30) | 46.9   | NR           | German              |
|    |                       |      |                |          |                |             |               |                 |                 | 224         | Inmates                                                         | 32.80 (9.60)  | 71.9   | NR           |                     |
|    |                       |      |                |          |                |             |               |                 |                 | 295         | University students                                             | 23.70 (2.80)  | 32.5   | NR           |                     |
| 12 | Cammack et al.        | 2016 | Epidemiologist | Yes      | English        | USA         | North America | Longitudinal    | Convenience     | 247         | Pregnant women involved in mental health program                | 32.70 (4.50)  | 0      | 89.1         | Original            |
| 13 | Charak et al.         | 2017 | Psychologist   | Yes      | English        | Burundi     | Africa        | Cross-sectional | Stratified      | 231         | Adolescents and youth                                           | 14.90 (1.99)  | 41.6   | NR           | Kirundi             |

(continued on next page)

Table 1 (continued)

|    | Authors           | Year | Affiliation*                 | Financed | Study language | Location     | Continent     | Research design | Sampling method   | Sample size | Population target                                    | Mean Age (SD) | % male | % Caucasians | Language adaptation |
|----|-------------------|------|------------------------------|----------|----------------|--------------|---------------|-----------------|-------------------|-------------|------------------------------------------------------|---------------|--------|--------------|---------------------|
| 14 | Jiang et al.      | 2018 | Psychiatrist                 | Yes      | English        | China        | Asia          | Longitudinal    | Convenience       | 200         | Schizophrenic patients                               | 28.30 (5.90)  | 50.0   | NR           | Chinese             |
| 15 | Sacchi et al.     | 2018 | Psychologist                 | No       | English        | Italy        | Europe        | Cross-sectional | Convenience       | 446         | University students                                  | 23.06 (2.87)  | 23.5   | NR           | Italian             |
| 16 | He et al.         | 2019 | Medical-Psychological Center | Yes      | English        | China        | Asia          | Cross-sectional | Incidental        | 3431        | University students                                  | 19.57 (1.03)  | 56.6   | NR           | Chinese             |
|    |                   |      |                              |          |                |              |               |                 |                   | 217         | Depressive clinical patients                         | 22.60 (5.70)  | 54.4   | NR           |                     |
| 17 | Kongerslev et al. | 2019 | Psychologist                 | No       | English        | Denmark      | Europe        | Cross-sectional | Archival material | 142         | Patients with personality disorders                  | 29.00 (8.40)  | 32.0   | NR           | Danish              |
|    |                   |      |                              |          |                |              |               |                 |                   | 101         | Patients with psychosis                              | 23.00 (3.40)  | 74.0   | NR           |                     |
|    |                   |      |                              |          |                |              |               |                 |                   | 71          | Patients with depression                             | 35.00 (11.60) | 22.0   | NR           |                     |
|    |                   |      |                              |          |                |              |               |                 |                   | 80          | Juvenile delinquents                                 | 17.00 (0.80)  | 100.0  | NR           |                     |
| 18 | Spies et al       | 2019 | Psychiatrist                 | Yes      | English        | South Africa | Africa        | Longitudinal    | Not mentioned     | 314         | African women with and without HIV                   | 30.70 (8.00)  | 0      | 0            | African             |
| 19 | Aloba et al.      | 2020 | Psychologist                 | No       | English        | Nigeria      | Africa        | Cross-sectional | Stratified        | 1337        | Adolescents                                          | 15.24 (1.43)  | 45.3   | 0            | Original            |
| 20 | Mizuki & Fujiwara | 2021 | Doctor                       | Yes      | English        | Japan        | Asia          | Cross-sectional | Snowball          | 31          | Institutionalized adolescents for child maltreatment | 15.80 (1.71)  | 61.3   | NR           | Japanese            |
|    |                   |      |                              |          |                |              |               |                 |                   | 46          | Community sample                                     | 15.60 (1.84)  | 54.4   | NR           |                     |
| 21 | Petrikova et al.  | 2021 | Social Health Institute      | Yes      | English        | Slovakia     | Europe        | Cross-sectional | Not mentioned     | 1018        | Representative Slovak sample                         | 46.20 (NR)    | 48.7   | NR           | Slovak              |
| 22 | Xiang et al.      | 2021 | Psychiatrist                 | Yes      | English        | China        | Asia          | Longitudinal    | Convenience       | 35          | Patients with schizophrenia                          | 17.91 (3.60)  | 42.9   | NR           | Chinese             |
| 23 | Hagborg et al.    | 2022 | Psychologist                 | Yes      | English        | Sweden       | Europe        | Longitudinal    | Not mentioned     | 1428        | Community adolescents W2                             | NR            | 48.2   | NR           | Swedish             |
|    |                   |      |                              |          |                |              |               |                 |                   | 1302        | Community adolescents W3                             | NR            | 49.5   | NR           |                     |
|    |                   |      |                              |          |                |              |               |                 |                   | 711         | Community adolescents W4                             | 15 (NR)       | 51.3   | NR           |                     |
|    |                   |      |                              |          |                |              |               |                 |                   | 935         | Community adolescents W5                             | 17(NR)        | 44.4   | NR           |                     |
|    |                   |      |                              |          |                |              |               |                 |                   | 74          | Adolescents clinical (misuse substance treatment)    | 18 (NR)       | 60.8   | NR           |                     |
| 24 | Wang et al.       | 2022 | Medical-Psychological Center | Yes      | English        | China        | Asia          | Cross-sectional | Convenience       | 1507        | Non-clinical adolescents                             | 16.10 (0.84)  | 54.8   | NR           | Chinese             |
|    |                   |      |                              |          |                |              |               |                 |                   | 281         | Adolescents with depression                          | 16.12 (0.82)  | 77.6   | NR           |                     |
| 25 | Cruz              | 2023 | Psychologist                 | No       | English        | USA          | North America | Cross-sectional | Not mentioned     | 863         | middle-aged Americans                                | 52.73 (13.41) | 48.0   | 70.0         | Original            |

\* Affiliation of corresponding author; SD = Standard Deviation. NR: Not Reported.

Measures (PROMs) (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018). The study of Georgieva et al. (2021) examined ten validations of the instrument and found that in eight of these, internal consistency was lower than 0.70 for at least one subscale, indicating consistent issues with physical neglect, as well as occasional problems with emotional neglect, emotional abuse, and physical abuse.

Using a Reliability Generalization Meta-analysis for the CTQ-SF is a promising approach to understanding its reliability across different samples and study features. Reliability generalization enables an examination of how reliability estimates of test scores vary across studies and samples, identifying associations with different studies and sample characteristics (Sánchez-Meca et al., 2013). Given that prior systematic reviews indicated concerns about the internal consistency of the CTQ-SF, this study aims to conduct a reliability generalization meta-analysis to enhance understanding of its reliability estimates. Notably, despite the widespread use of CTQ-SF in assessing childhood maltreatment, no previous study has undertaken a reliability generalization analysis on this instrument (Madison et al., 2020). Therefore, it is crucial to examine the reliability of studies that use this instrument across a range of samples and study characteristics. Consequently, this study has the following aims: (1) to estimate the average reliability of test scores (for the total scale and factors), in terms of internal consistency and test–retest reliability, found in empirical studies that applied it and reported reliability estimates with the data at hand; (2) to analyze the heterogeneity among the reliability estimates; and (3) to search for study characteristics that can be statistically related to the reliability estimates.

## 2. Methods

To ensure the transparency and accuracy of the systematic review, this study adhered to the Preferred Reporting Items for Systematic Review and Meta-analysis Protocols (PRISMA-P) guidelines developed by Moher et al. (2015) and was conducted according to the recommendations for conducting and reporting Reliability Generalization Meta-Analyses (REGEMA Checklist; Sánchez-Meca et al., 2021).

### 2.1. Eligibility criteria

The present study employed a comprehensive selection process to identify articles that directly analyzed the psychometric properties of the CTQ-SF, as developed by Bernstein et al. (2003). The inclusion criteria were designed to ensure the thoroughness and rigor of the data, and thus, studies that met the following conditions were considered: 1) original and quantitative research published in English, Spanish, Catalanian, Bulgarian, or French in peer-reviewed journals, as these are the languages that researchers dominate; 2) studies published between January 2003 and December 2022, in order to include all studies from the development of the instrument; 3) articles published in scientific journals indexed in the Journal Citation Reports (JCR) or Scopus, as assessed by Scimago Journal and Country Rank, to ensure quality and peer-reviewed research; 4) studies focused in examining the psychometric properties of the CTQ-SF or any of its adaptations to other languages and cultures while maintaining the 28 or 25-item structure (without the 3 minimization and denial items) and the 5-point Likert-type scale; as all studies utilizing the CTQ-SF yielded to an unbearable number of articles; 5) use of any type of target population, and; 6) to report reliability estimates based on the study-specific sample for total scores (e.g., Cronbach's alpha coefficient, omega coefficient, and/or test-retest reliability). Additionally, a minimum sample size for the inclusion of a study was not established. However, the quality of all included studies was assessed and included as supplementary materials.

Given the retrospective nature of the CTQ-SF, which is designed to assess experiences during childhood and adolescence, no limits were set on subject ages. However, the exclusion criteria were established to maintain a high level of scientific integrity, and thus, articles considered gray literature (such as doctoral theses, conference papers, and press articles) and articles presenting results about the CTQ without psychometric information were not included. Also, case series studies and studies that used another version without the 25-item CTQ-SF structure were excluded. This decision was taken due to the impracticality of conducting reliable comparisons with other versions, especially when there is only one study providing information about it. A single adaptation lacks the substantive basis required for performing comprehensive reliability analysis, rendering it insufficient for meaningful cross-version assessments and comparisons.

### 2.2. Search strategy

A systematic search of the electronic databases of ProQuest, Scopus, and Web of Science was performed. All articles in English, Spanish, Catalanian, Bulgarian, and French were included. Two groups of keywords were identified, the complete name of the instrument "Childhood Trauma Questionnaire Short Form" and its shortened name "CTQ-SF". The search was performed during the months of January and February in 2023. Additionally, a second search was performed in July in order to update all published studies up to that date.

In addition, the lists of references from previous systematic reviews (Georgieva et al., 2021; Georgieva et al., 2022; Saini et al., 2019) and from previous studies that examined the psychometric properties of the CTQ-SF were screened to find studies that met the inclusion criteria for the present meta-analysis. Lastly, on the 9th of July 2023, a final search was performed using Google Scholar and other databases to locate all the studies that cited the study by Bernstein et al. (2003).

### 2.3. Selection process

The identification and selection process of eligible studies was conducted by two independent researchers, who screened the titles and abstracts obtained from the search strategy, as shown in Fig. 1. Any discrepancies between the researchers were discussed with a third researcher, as recommended by Bekhet and Zauszniewski (2012).

The results of the search were transferred into an electronic reference manager (Zotero 5.0 for Windows), which eliminated any duplicate entries. From a total of 14,113 studies identified, 26 met the inclusion criteria and were selected for this review, as illustrated in Fig. 1. Thus, the present Reliability Generalization (RG) study focused on the 26 studies that provided reliability estimates. Moreover, as several studies provided reliability estimates for two or more different samples, the database of the present RG meta-analysis included a total of 39 independent samples, which finally provided a total of 243 reliability coefficients (Cronbach's alpha, McDonald's Omega, and/or test-retest correlations).

### 2.4. Data collection process

Data collection was conducted by two independent researchers, who extracted the characteristics of the studies in a standardized and systematic manner by applying the established protocol. Information extracted from eligible studies were as follows: (1) name of the article, (2) authors, (3) year of publication, (4) first authors' affiliation, (5) financing of the study, (6) language of the article, (7) location and continent where the investigation was carried out, (8) research design, (9) type of sampling, (10) total sample size, (11) type of population (considering students, clinical population, substance users, etcetera), (12) mean and standard deviation (SD) of the participants' age (in years), (13) male percentage, (14) Caucasian percentage, (15) if there was a language adaptation carried out, and to which language, (16) mean and SD of the total scores of the scale and each subscale independently, (17) Cronbach's alpha and/or McDonald's omega of the total scale and for each factor, (18) if test-retest analysis were carried out, results of the test-retest analysis, sample size of the test-retest analysis and the time interval (in weeks) between the measures.

### 2.5. Reliability estimates

Two types of reliability coefficients were considered: internal consistency (e.g., Cronbach's alpha coefficient and McDonald's omega coefficient) and test-retest reliability (Pearson correlation coefficient). However, inter-rater reliability was not considered. Both types of reliability coefficients were extracted for the CTQ-SF total score and for each of the five subscales. Alpha and omega are based on different assumptions (Yang & Green, 2011). Alpha is a biased estimate unless items are unidimensional and tau-equivalent, it is based on the assumptions of normal distribution, equal variance, and additionally, alpha depends on the variability in the population (sample) analyzed (Viladrich et al., 2017). On the contrary, omega is more robust than alpha against deviations from these assumptions and therefore, a more suitable measure of internal consistency, as it is based on factor analytic techniques that can accommodate congeneric and non-normal data (Gabler & Raykov, 2017). The reliability coefficients were transformed in order to normalize their distributions and stabilize their variances. Cronbach's alpha coefficients and McDonald's omega coefficients were transformed by applying the formula proposed by Bonett (2002), and Pearson correlation coefficients,  $r$ , were transformed to Fisher's  $z$  (Sánchez-Meca et al., 2013). Afterwards, to facilitate the interpretation of the results of each meta-analysis, the average and individual reliability coefficients, along with their confidence limits, obtained with Bonett's or Fisher's  $z$  transformations, were back-transformed into the Cronbach's alpha coefficient and McDonald's omega coefficient and Pearson correlation metrics, respectively.

### 2.6. Methodological quality of included studies

Risk of bias and the methodological quality of all the studies included was assessed by two independent researchers in order to determine the validity of the results. All discrepancies were discussed with a third researcher. The COSMIN protocol for assessing validations scientific quality (Mokkink et al., 2018; Prinsen et al., 2018; Terwee et al., 2018) was used.

The initial section of this tool comprises three primary categories: content validity, internal structure, and remaining measurement properties. Content validity, despite PROM development not being categorized as a measurement property, was delineated into two sections on the instrument: PROM development and content validity. The authors emphasized that although PROM development doesn't fall under a measurement property, it remains pertinent to the assessment of content validity. Internal structure consisted of three segments: structural validity, internal consistency, and cross-cultural validity/measurement invariance. Lastly, the remaining measurement properties—reliability, measurement error, criterion validity, hypothesis testing for construct validity, and responsiveness—were appraised through individual boxes. It is important to note that the COSMIN taxonomy utilizes PROM as a broader term that also includes instruments used to assess a variety of constructs.

All the criteria were rated on a 4-point scale, spanning from excellent to poor. The overall score for each property was determined by the lowest score (1–4) assigned to an item within each box. The assessment scores for each individual study are provided in a separate table included as supplementary material. One study was excluded at this point (Behn et al., 2020) for not reporting any reliability information.

### 2.7. Statistical analysis

Separate RG meta-analyses were conducted for Cronbach's alpha coefficients, McDonald's omega coefficients, and Pearson's

correlation coefficients (test-retest reliability). Moreover, independent RG meta-analyses were performed for the reliability coefficients obtained for the CTQ-SF total scale scores and for each of its five subscales. In these statistical analyses, random-effects models were applied, and the coefficients were weighted by inverse variance (Borenstein et al., 2009; López-López et al., 2013; Sánchez-Meca et al., 2013). In addition, the restricted maximum likelihood was used to estimate between-study variance ( $\tau^2$ ). In each RG meta-analysis, an average reliability coefficient and a 95 % confidence interval were computed with the improved method proposed by Hartung and Knapp (2001) (see also Sánchez-Meca & Marín-Martínez, 2008), and a forest plot was constructed to display these reliability estimates and to visually analyze the heterogeneity among reliability estimates in each meta-analysis. Both additions, Cochran's  $Q$ -statistic and the  $I^2$  index were calculated to evaluate the variability among the reliability estimates. A  $Q$ -statistic with  $p < .05$  was indicative of heterogeneity among the reliability estimates (Higgins et al., 2003).  $I^2$  values of around 25 %, 50 %, and 75 % denoted low, moderate, and large heterogeneity, respectively (Huedo-Medina et al., 2006). Additionally, sensitivity analyses were conducted to examine the robustness of the results and the impact of including or excluding one-to-one of reliability coefficients.

Lastly, for meta-analyses with at least 10 reliability estimates moderator analyses were conducted to detect the study characteristics related to reliability and whether the reliability estimates exhibited heterogeneity. For categorical moderators (such as study design, financial source, geographical location, population target, researcher affiliation, and study language) Weighted ANOVAs were performed while for continuous moderators (such as study mean and standard deviation of scores, sample size, percentage of male, participants' mean age, and year publication of study) meta-regressions using the improved F-statistic proposed by Knapp and Hartung (2003) were estimated. In all cases, a mixed-effects model was assumed (López-López et al., 2013; Viechtbauer et al., 2015). The proportion of variance accounted for by each moderator variable was estimated with the  $R^2$  index (López-López et al., 2014; Raudenbush, 2009). All statistical analyses were conducted using the metafor package in R version 4.2.1 (Viechtbauer, 2010).

### 3. Results

#### 3.1. Characteristics of studies

Table 1 presents the characteristics of the included studies. There was a slight increase in publications during the years 2019–2021 with three validations per year, while the rest of the years ranged between one and two validations each year. The most common affiliation of the corresponding author was a psychologist (44 %, 11 of 25 studies), followed by psychiatrist (28 %, 7 of 26 studies) although the instrument has also been validated by criminologists, doctors, and social workers among others. The majority of studies were financed (68 %, 17 of 26 studies).

Regarding language and location, all studies were published in English. The most common continent in which the validations were performed was Europe (48 %, 12 of 25 studies) followed by Asia (24 %, 6 of 25 studies) and North America (16 %, 4 of 25 studies); however, if we examine the specific countries, the country in which this instrument has been validated the most times if the United States (16 %, 4 of 26 studies) and China (16 %, 4 of 25 studies), followed by Netherlands (12 %, 3 of 25 studies) and Sweden (8 %, 2 of 26 studies).

In addition, the most common research design was cross-sectional (72 %, 18 of 25 studies) although there was also a substantial portion of longitudinal studies (28 %, 7 of 25 studies). The sampling method was not mentioned in many studies (44 %, 11 of 26 studies) while in the studies that were mentioned, the most common method was convenience sampling (28 %, 7 of 25 studies) followed by stratified sampling (12 %, 3 of 25 studies).

Participants in the studies exhibited a wide range of characteristics and diverse origins. The average sample size was 554 participants ( $SD = 691$ ), ranging from 35 to 3431. The mean age of all included samples was 28.91 years ( $SD = 10.45$ ,  $Min. = 14.9$ ,  $Max. = 52.73$ ). The average percentage of males in all studies was 44.5 %, and the average Caucasian percentage was 48.4. Finally, the most common language in which the instrument was validated was its original language (English) although we could identify a total of 13 other languages of validation.

Regarding CTQ-SF total scores, for all samples, the average of the mean CTQ-SF total scores was 46.05 ( $SD = 9.68$ ,  $Min. = 31.15$ ,  $Max. = 65.10$ ) and the average of the standard deviation for CTQ-SF total scores was 13.35 ( $SD = 4.30$ ,  $Min. = 5.90$ ,  $Max. = 19.84$ ). For the physical abuse subscale, the average of the mean total scores was 7.42 ( $SD = 1.80$ ,  $Min. = 5.50$ ,  $Max. = 12.80$ ) and the average of the standard deviation for total scores was 3.42 ( $SD = 1.40$ ,  $Min. = 1.63$ ,  $Max. = 5.85$ ), for the emotional abuse subscale, the average of the mean total scores was 9.86 ( $SD = 2.75$ ,  $Min. = 6.30$ ,  $Max. = 18.0$ ) and the average of the standard deviation for total scores was 4.54 ( $SD = 1.35$ ,  $Min. = 1.90$ ,  $Max. = 7.72$ ), for the sexual abuse subscale the average of the mean total scores was 6.74 ( $SD = 1.34$ ,  $Min. = 5.08$ ,  $Max. = 11.30$ ) and the average of the standard deviation for total scores was 3.42 ( $SD = 1.57$ ,  $Min. = 0.67$ ,  $Max. = 6.6$ ); for the physical neglect subscale the average of the mean total scores was 8.31 ( $SD = 1.90$ ,  $Min. = 4.55$ ,  $Max. = 12.60$ ) and the average of the standard deviation for total scores was 3.23 ( $SD = 0.92$ ,  $Min. = 1.51$ ,  $Max. = 5.25$ ); and the for emotional neglect subscale the average of the mean total scores was 12.12 ( $SD = 3.06$ ,  $Min. = 7.35$ ,  $Max. = 18.86$ ) and the average of the standard deviation for total scores was 5.15 ( $SD = 1.83$ ,  $Min. = 3.10$ ,  $Max. = 12.10$ ).

#### 3.2. Average reliability and heterogeneity

Separate RG meta-analyses were conducted for the CTQ-SF total score reliability and for each of the factors. Moreover, different RG meta-analyses were performed for Cronbach's alpha coefficient, McDonald's Omega coefficient, and test-retest reliability coefficients. Table 2 displays the average reliability for each type of coefficient and the 95 % confidence interval, alongside the  $Q$  test and the  $I^2$  index for the CTQ-SF total Score and its five subscales. It can be noted that Cronbach's alpha coefficient was the most prevalently

reported. The eighteen Cronbach's alpha coefficient reported for the CTQ-SF total scores ranged from 0.79 to 0.94, with an average Cronbach's alpha coefficient of 0.891 (95 % CI: 0.868, 0.910) (see Fig. 2 in Appendix). For the physical abuse subscale, the Cronbach's coefficient estimates ranged from 0.50 to 0.95, with an average reliability of 0.843 (95 % CI: 0.818, 0.865,  $k = 36$ ) (see Fig. 3 in Appendix); for the emotional abuse subscale the reliability estimates ranged from 0.67 to 0.91 with an average Cronbach's alpha coefficient of 0.839 (95 % CI: 0.819, 0.857,  $k = 36$ ) (see Fig. 4 in Appendix); for the sexual abuse subscale, the Cronbach's alpha coefficient ranged from 0.72 to 0.97, with an average reliability estimate of 0.916 (95 % CI: 0.896, 0.931,  $k = 37$ ) (see Fig. 5 in Appendix), for the physical neglect subscale the reliability estimates ranged from 0.36 to 0.87, with an average alpha coefficient of 0.656 (95 % CI: 0.606, 0.698,  $k = 36$ ) (see Fig. 6 in Appendix), and for the emotional neglect subscale reliability estimates ranged from 0.54 to 0.92 with an average Cronbach's alpha coefficient of 0.851 (95 % CI: 0.827, 0.872,  $k = 35$ ) (see Fig. 7 in Appendix). In all cases, there was evidence of large heterogeneity among Cronbach's coefficient estimates, with all  $Q$  statistics being statistically significant and  $I^2$  indexes ranging from 96.4 % to 98.6 %, all indicative of large magnitude.

Regarding the test-retest reliability, the four estimates reported for the CTQ-SF total score ranged from 0.72 to 0.87 with an average test-retest coefficient of 0.788 (95 % CI 0.658, 0.872) and the average test-retest estimates for the five subscales ranged between 0.668 and 0.709. The average time interval between test-retest administrations was of 39.15 weeks ( $SD = 59.58$ ,  $Min. = 2$ ,  $Max. = 141$ ). Again, there was evidence of large heterogeneity among test-retest estimates, with all  $Q$  statistics being statistically significant and  $I^2$  indexes ranging from 66.1 % to 94.0 %, all indicative of large magnitude.

### 3.3. Sensitivity analyses

Sensitivity analyses were performed to examine the robustness of the results and the impact of including or excluding one-to-one of reliability coefficients. By deleting each study one by one and performing the analyses again each time, the average Cronbach's alpha coefficient ranged from 0.887 (95 % CI: 0.863, 0.906) to 0.895 (95 % CI: 0.874, 0.913) for CTQ-SF total scores. For subscales, the average Cronbach's alpha coefficient ranged from 0.838 (95 % CI: 0.814, 0.858) to 0.848 (95 % CI: 0.825, 0.867) for physical abuse; from 0.836 (95 % CI: 0.816, 0.854) to 0.842 (95 % CI: 0.823, 0.859) for emotional abuse; from 0.913 (95 % CI: 0.894, 0.929) to 0.918 (95 % CI: 0.900, 0.933) for sexual abuse; from 0.644 (95 % CI: 0.598, 0.685) to 0.661 (95 % CI: 0.614, 0.703) for physical neglect; and from 0.849 (95 % CI: 0.824, 0.870) to 0.856 (95 % CI: 0.835, 0.875) for emotional neglect. These results indicate that the average reliability estimate is quite robust.

Concerning the McDonald's Omega coefficient, the average estimates ranged from 0.800 (95 % CI: 0.781, 0.817) to 0.800 (95 % CI: 0.784, 0.815) for CTQ-SF total scores. For subscales, the average McDonald's Omega coefficients ranged from 0.828 (95 % CI: 0.000, 1) to 0.940 (95 % CI: 0.000, 0.1) for physical abuse; from 0.784 (95 % CI: 0.000, 0.1) to 0.898 (95 % CI: 0.000, 0.995) for emotional abuse; from 0.750 (95 % CI: 0.730, 0.769) to 0.960 (95 % CI: 0.956, 0.963) for sexual abuse; from 0.678 (95 % CI: 0.000, 0.997) to 0.807 (95 % CI: 0.007, 0.962) for physical neglect; and from 0.861 (95 % CI: 0.549, 0.957) to 0.938 (95 % CI: 0.276, 0.995) for emotional neglect. These results indicate that the average reliability estimate is quite robust.

**Table 2.**

Average Reliability Coefficients, 95 % confidence intervals, and heterogeneity statistics for the CTQ-SF Total score and its five subscales.

|                                     | $k$ | $r_+$ | 95 % CI |       | $Q$         | $I^2$ | $\tau^2_{REML}$ |
|-------------------------------------|-----|-------|---------|-------|-------------|-------|-----------------|
|                                     |     |       | LL      | UL    |             |       |                 |
| <b>Cronbach's Alpha coefficient</b> |     |       |         |       |             |       |                 |
| Total score                         | 18  | 0.891 | 0.868   | 0.910 | 686.227***  | 96.8  | 0.136           |
| Physical abuse                      | 36  | 0.843 | 0.818   | 0.865 | 1255.572*** | 97.1  | 0.179           |
| Emotional abuse                     | 36  | 0.839 | 0.819   | 0.857 | 1315.404*** | 95.4  | 0.112           |
| Sexual abuse                        | 37  | 0.916 | 0.896   | 0.931 | 3380.408*** | 98.5  | 0.363           |
| Physical neglect                    | 36  | 0.656 | 0.606   | 0.698 | 798.414***  | 96.2  | 0.137           |
| Emotional neglect                   | 35  | 0.851 | 0.827   | 0.872 | 1205.275*** | 97.2  | 0.185           |
| <b>McDonald's Omega coefficient</b> |     |       |         |       |             |       |                 |
| Total score                         | 2   | 0.800 | 0.800   | 0.800 | 0.000       | 0.0   | 0.000           |
| Physical abuse                      | 3   | 0.894 | 0.000   | 0.993 | 1806.584*** | 99.9  | 1.132           |
| Emotional abuse                     | 3   | 0.845 | 0.000   | 0.977 | 919.229***  | 99.8  | 0.588           |
| Sexual abuse                        | 2   | 0.900 | 0.000   | 1.000 | 930.015***  | 99.9  | 1.677           |
| Physical neglect                    | 3   | 0.740 | 0.033   | 0.930 | 424.521***  | 99.5  | 0.278           |
| Emotional neglect                   | 4   | 0.914 | 0.549   | 0.984 | 2814.796*** | 99.8  | 1.089           |
| <b>Test-retest coefficient</b>      |     |       |         |       |             |       |                 |
| Total score                         | 4   | 0.788 | 0.658   | 0.872 | 13.506**    | 75.6  | 0.021           |
| Physical abuse                      | 5   | 0.709 | 0.461   | 0.854 | 31.980***   | 94.0  | 0.080           |
| Emotional abuse                     | 5   | 0.704 | 0.624   | 0.769 | 13.827**    | 66.1  | 0.010           |
| Sexual abuse                        | 4   | 0.677 | 0.216   | 0.891 | 34.179***   | 94.9  | 0.126           |
| Physical neglect                    | 5   | 0.668 | 0.416   | 0.824 | 54.109***   | 93.8  | 0.078           |
| Emotional neglect                   | 5   | 0.681 | 0.545   | 0.783 | 12.730*     | 74.1  | 0.015           |

Note.  $k$  = number of reliability estimates.  $r_+$  = average reliability coefficient. LL and UL = lower and upper limits of the 95 % confidence interval for  $r_+$ .  $Q$  = Cochran's heterogeneity  $Q$  statistic;  $Q$  statistic has  $k - 1$  degrees of freedom.  $I^2$  = heterogeneity index;  $\tau_{REML}^2$  = between-studies variance estimated using restricted maximum likelihood. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

Regarding the test-retest reliability, the average test-retest estimates ranged from 0.765 (95 % CI: 0.605, 0.865) to 0.816 (95 % CI: 0.712, 0.884) for CTQ-SF total scores. For subscales, the average test-retest estimates ranged from 0.633 (95 % CI: 0.429, 0.776) to 0.747 (95 % CI: 0.446, 0.896) for physical abuse; from 0.683 (95 % CI: 0.574, 0.769) to 0.732 (95 % CI: 0.684, 0.774) for emotional abuse; from 0.565 (95 % CI: 0.131, 0.817) to 0.742 (95 % CI: 0.073, 0.950) for sexual abuse; from 0.629 (95 % CI: 0.257, 0.839) to 0.721 (95 % CI: 0.540, 0.839) for physical neglect; and from 0.662 (95 % CI: 0.549, 0.751) to 0.701 (95 % CI: 0.631, 0.760) for emotional neglect. These results indicate that the average reliability estimate is quite robust.

### 3.4. Moderator analyses

Table 3 displays the results of the simple meta-regression analyses for the continuous moderator variables applied to Cronbach's alpha coefficients for the CTQ-SF total score and its five subscales. It can be noted that the variability of total scores in CTQ-SF ( $p = .012$ ,  $R^2 = 0.524$ ), physical abuse ( $p = .027$ ,  $R^2 = 0.190$ ), emotional abuse ( $p < .001$ ,  $R^2 = 0.491$ ), sexual abuse ( $p = .047$ ,  $R^2 = 0.154$ ), and physical neglect ( $p = .043$ ,  $R^2 = 0.156$ ) showed a statistically significant association with Cronbach's alpha coefficients, such that the larger the variability observed, the larger the Cronbach's alpha coefficient. None of the moderators examined exhibited a relationship with Cronbach's alpha coefficients for the emotional neglect subscale.

Furthermore, regarding the overall CTQ-SF score, there was a statistically significant association between sample size ( $p = .049$ ,  $R^2 = 0.195$ ) and Cronbach's alpha coefficients. To identify the most important study characteristics to explain the variability of the Cronbach's alpha coefficients for the CTQ-SF total score, a multiple meta-regression was performed. The full model performed on 10 studies showed a marginally significant association with Cronbach's Alpha coefficients,  $F(2, 7) = 4.655$ ,  $p = .052$ , with 45.0 % of the variance explained. Of the two model predictors, only the variability of scores showed a marginally significant association with Cronbach's alpha coefficients after controlling for the influence of the sample size ( $p = .0509$ ), suggesting that the Cronbach's Alpha coefficients obtained in the studies were higher when the variability of scores observed was larger.

Concerning to emotional abuse subscale, in addition to the variability of scores, the sample size showed a negative association with Cronbach's alpha coefficients ( $p = .013$ ,  $R^2 = 0.152$ ) as well as participants' mean age participants' mean age ( $p = .049$ ,  $R^2 = 0.107$ ). The variability of participants' age showed a positive association ( $p = .013$ ,  $R^2 = 0.189$ ), and the year of publication of the study, which showed a negative association ( $p = .026$ ,  $R^2 = 0.143$ ). A multiple meta-regression was performed to detect the most relevant study characteristics to explain the variability of Cronbach's alpha coefficients (see Table 4). The full model exhibited a statistically significant relationship with Cronbach's Alpha coefficients,  $F(6, 13) = 4.369$ ,  $p = .012$ , with 54.0 % of the variance explained. Of the six model predictors, only the variability of scores exhibited a statistically significant relationship with Cronbach's alpha coefficients after controlling for the influence of the other variables ( $p = .039$ ), thus, the Cronbach's alpha coefficients obtained in the studies were higher when the variability of scores observed was larger.

Regarding physical neglect (in addition to the variability of scores), the mean total score showed a positive association ( $p = .042$ ,  $R^2 = 0.151$ ) and exhibited a statistically significant relationship with Cronbach's alpha coefficients. Again, to identify the most important study characteristics to explain the variability of Cronbach's alpha coefficients for physical neglect, a multiple meta-regression was performed. The full model did not show a statistically significant association with Cronbach's Alpha coefficients ( $F(1, 19) = 2.388$ ,  $p = .0119$ ,  $R^2 = 0.121$ ,  $k = 22$ ).

Table 5 displays the results of the ANOVAs for the categorical moderator variables performed on Cronbach's alpha coefficients for the CTQ-SF total scores and its five subscales. Regarding the CTQ-SF total scores, only one categorical moderator variable showed a statistically significant association with the Cronbach's Alpha estimates: the geographical location where the study was conducted ( $p = .007$ ). Specifically, the average reliability was lower for studies conducted in Asian countries ( $\alpha_+ = 0.848$ ), compared to studies conducted in Europe ( $\alpha_+ = 0.910$ ) and North America ( $\alpha_+ = 0.915$ ), explaining 46.4 % of the variance. In addition, it is worth noting that the affiliation of the main researcher of the study showed no statistically significant relation with the reliability estimates ( $p = .054$ ), but it accounted for a large proportion of variance (40.7 %) for the CTQ-SF total score. The average reliability was lower for studies conducted by people who belonged to the Medical Psychological Center ( $\alpha_+ = 0.830$ ), compared to studies conducted by other professionals.

With regard to the emotional abuse and sexual abuse subscales, three of the categorical moderator variables showed a statistically significant relationship with the Cronbach's alpha coefficient for the total scores: the financial source ( $p = .003$ ,  $p = .047$ , respectively), the affiliation of the main researcher of the study ( $p = .042$ ,  $p = .027$ , respectively), and the geographical location where the study was conducted ( $p = .003$ ,  $p = .005$ , respectively). The financial sources, classified as yes versus no/not reported, accounted for 24.7 % of the variance for emotional abuse and 9.1 % of the variance for Sexual abuse. In both cases, the average reliability was higher for studies that were not financed ( $\alpha_+ = 0.864$ ,  $\alpha_+ = 0.932$ , respectively), compared to studies that received funding ( $\alpha_+ = 0.811$ ,  $\alpha_+ = 0.899$ , respectively). In addition, the affiliation of the main researcher of the study accounted for 20.5 % of the variance for emotional abuse and 22.6 % of the variance for sexual abuse. In both cases, the average reliability was lower for studies conducted by people who belonged to Medical Psychological Centers ( $\alpha_+ = 0.735$ ,  $\alpha_+ = 0.797$ , respectively), compared to studies conducted by other professionals. Moreover, the geographical location of the study accounted for 36.1 % of the variance for emotional abuse and 31.6 % of the variance for sexual abuse. In both cases, the average reliability was lower for studies conducted in Africa ( $\alpha_+ = 0.718$ ,  $\alpha_+ = 0.852$ , respectively) and Asian countries ( $\alpha_+ = 0.784$ ,  $\alpha_+ = 0.844$ , respectively), compared to studies conducted in Europe ( $\alpha_+ = 0.858$ ,  $\alpha_+ = 0.935$ , respectively), North America ( $\alpha_+ = 0.857$ ,  $\alpha_+ = 0.926$ , respectively) and South America ( $\alpha_+ = 0.850$ ,  $\alpha_+ = 0.930$ , respectively).

Regarding emotional neglect subscale, three of the categorical moderator variables revealed a statistically significant association with the Cronbach's alpha coefficients for the total scores: the type of research design ( $p = .023$ ), the financial source ( $p < .001$ ), and the affiliation of the main researcher of the study ( $p = .004$ ). The type of research design, classified as longitudinal vs. cross-sectional,

accounted for 12.0 % of the variance for emotional neglect. Specifically, the average reliability was better for studies that used a cross-sectional design ( $\alpha_+ = 0.861$ ), than those that used a longitudinal design ( $\alpha_+ = 0.774$ ). The financial sources accounted for 44.2 % of the variance for emotional neglect, with the largest average Cronbach's alpha coefficients for studies that were not financed ( $\alpha_+ = 0.891$ ), compared to studies that received funding ( $\alpha_+ = 0.806$ ). Moreover, the affiliation of the main researcher of the study accounted for 35.1 % of the variance for emotional neglect, with the smallest average Cronbach's alpha coefficients for studies conducted by criminologists ( $\alpha_+ = 0.655$ ), compared to studies conducted by other professionals.

Lastly, concerning the physical abuse and physical neglect subscales, none of the moderator variables examined exhibited a statistically significant relationship with Cronbach's alpha coefficient for the total scores on the physical abuse and physical neglect

**Table 3.**

Results of the simple meta-regressions applied on Cronbach's alpha coefficients for the CTQ-SF total score and its five subscales, taking continuous moderator variables as predictors.

| Predictor variable | k  | b <sub>j</sub> | SE     | 95 % CI |         | F       | p      | Q <sub>E</sub> | R <sup>2</sup> |
|--------------------|----|----------------|--------|---------|---------|---------|--------|----------------|----------------|
|                    |    |                |        | LL      | UL      |         |        |                |                |
| CTQ-SF TOTAL       |    |                |        |         |         |         |        |                |                |
| Mean total score   | 10 | −0.013         | 0.013  | −0.044  | 0.018   | 0.951   | 0.358  | 127.291***     | 0.000          |
| SD of total score  | 10 | 0.085          | 0.026  | 0.024   | 0.145   | 10.515  | 0.012  | 113.479***     | 0.524          |
| N                  | 18 | −0.0002        | 0.0001 | −0.0004 | 0.0000  | 4.527   | 0.049  | 239.094***     | 0.195          |
| % male             | 18 | −0.004         | 0.004  | −0.012  | 0.004   | 1.246   | 0.280  | 637.146***     | 0.015          |
| Mean age (years)   | 17 | 0.017          | 0.0091 | −0.002  | 0.037   | 3.707   | 0.073  | 325.238***     | 0.163          |
| SD of age (years)  | 15 | 0.051          | 0.026  | −0.004  | 0.107   | 4.144   | 0.063  | 239.938***     | 0.196          |
| Year               | 18 | −0.025         | 0.019  | −0.066  | 0.016   | 1.624   | 0.221  | 644.321***     | 0.047          |
| PHYSICAL ABUSE     |    |                |        |         |         |         |        |                |                |
| Mean total score   | 23 | 0.066          | 0.059  | −0.057  | 0.188   | 1.244   | 0.277  | 845.912***     | 0.010          |
| SD of total score  | 22 | 0.169          | 0.071  | 0.021   | 0.317   | 5.703   | 0.027  | 637.606***     | 0.190          |
| N                  | 36 | −0.0001        | 0.0001 | −0.0004 | 0.0001  | 1.209   | 0.279  | 871.105***     | 0.009          |
| % male             | 35 | −0.005         | 0.003  | −0.011  | 0.001   | 2.468   | 0.126  | 1174.482***    | 0.037          |
| Mean age (years)   | 35 | 0.010          | 0.007  | −0.005  | 0.025   | 1.740   | 0.196  | 1022.162***    | 0.023          |
| SD of age (years)  | 31 | 0.033          | 0.018  | −0.005  | 0.070   | 3.188   | 0.085  | 836.811***     | 0.074          |
| Year               | 36 | −0.011         | 0.012  | −0.036  | 0.014   | 0.819   | 0.372  | 1137.015***    | 0.000          |
| EMOTIONAL ABUSE    |    |                |        |         |         |         |        |                |                |
| Mean total score   | 23 | 0.024          | 0.027  | −0.033  | 0.081   | 0.740   | 0.400  | 579.827***     | 0.000          |
| SD of total score  | 22 | 0.191          | 0.044  | 0.098   | 0.284   | 19.0113 | <0.001 | 190.070***     | 0.491          |
| N                  | 36 | −0.0002        | 0.0001 | −0.0004 | −0.0000 | 6.862   | 0.013  | 774.855***     | 0.152          |
| % male             | 35 | −0.001         | 0.002  | −0.006  | 0.004   | 0.102   | 0.752  | 1269.380***    | 0.000          |
| Mean age (years)   | 35 | 0.011          | 0.005  | 0.0001  | 0.022   | 4.185   | 0.049  | 711.808***     | 0.107          |
| SD of age (years)  | 31 | 0.033          | 0.012  | 0.008   | 0.059   | 7.086   | 0.013  | 472.782***     | 0.189          |
| Year               | 36 | −0.021         | 0.009  | −0.040  | −0.003  | 5.454   | 0.026  | 883.031***     | 0.143          |
| SEXUAL ABUSE       |    |                |        |         |         |         |        |                |                |
| Mean total score   | 23 | 0.131          | 0.102  | −0.082  | 0.342   | 1.642   | 0.214  | 1696.091***    | 0.032          |
| SD of total score  | 22 | 0.178          | 0.084  | 0.002   | 0.353   | 4.462   | 0.047  | 1147.341***    | 0.154          |
| N                  | 37 | −0.0002        | 0.0002 | −0.0005 | 0.0002  | 1.233   | 0.275  | 2206.745***    | 0.007          |
| % male             | 36 | 0.004          | 0.004  | −0.004  | 0.012   | 0.971   | 0.331  | 3298.842***    | 0.000          |
| Mean age (years)   | 36 | 0.013          | 0.010  | −0.008  | 0.034   | 1.652   | 0.207  | 1939.836***    | 0.025          |
| SD of age (years)  | 32 | 0.036          | 0.026  | −0.017  | 0.089   | 1.997   | 0.170  | 1457.272***    | 0.039          |
| Year               | 37 | −0.025         | 0.017  | −0.060  | 0.009   | 2.170   | 0.150  | 2852.173***    | 0.036          |
| PHYSICAL NEGLECT   |    |                |        |         |         |         |        |                |                |
| Mean total score   | 23 | 0.106          | 0.049  | 0.004   | 0.208   | 4.674   | 0.042  | 547.206***     | 0.151          |
| SD of total score  | 22 | 0.215          | 0.099  | 0.008   | 0.421   | 4.696   | 0.043  | 530.959***     | 0.156          |
| N                  | 36 | −0.0002        | 0.0001 | −0.0004 | 0.0000  | 2.575   | 0.118  | 641.106***     | 0.050          |
| % male             | 35 | −0.001         | 0.003  | −0.006  | 0.005   | 0.080   | 0.780  | 771.143***     | 0.000          |
| Mean age (years)   | 35 | −0.005         | 0.006  | −0.018  | 0.008   | 0.636   | 0.431  | 656.616***     | 0.000          |
| SD of age (years)  | 31 | −0.010         | 0.017  | −0.045  | 0.024   | 0.379   | 0.543  | 631.171***     | 0.000          |
| Year               | 36 | −0.011         | 0.012  | −0.036  | 0.014   | 0.819   | 0.372  | 1137.015***    | 0.000          |
| EMOTIONAL NEGLECT  |    |                |        |         |         |         |        |                |                |
| Mean total score   | 22 | −0.039         | 0.035  | −0.113  | 0.034   | 1.260   | 0.275  | 706.561***     | 0.008          |
| SD of total score  | 22 | −0.096         | 0.059  | −0.219  | 0.028   | 2.614   | 0.122  | 1008.309***    | 0.080          |
| N                  | 35 | −0.0000        | 0.0001 | −0.0003 | 0.0002  | 0.143   | 0.707  | 961.470***     | 0.000          |
| % male             | 34 | 0.001          | 0.003  | −0.005  | 0.008   | 0.136   | 0.715  | 1099.013***    | 0.000          |
| Mean age (years)   | 34 | 0.013          | 0.007  | −0.002  | 0.028   | 2.924   | 0.097  | 900.146***     | 0.062          |
| SD of age (years)  | 30 | 0.025          | 0.019  | −0.014  | 0.064   | 1.763   | 0.195  | 803.962***     | 0.030          |
| Year               | 35 | −0.004         | 0.013  | −0.030  | 0.022   | 0.091   | 0.765  | 1158.482***    | 0.000          |

*Note.* SD = Standard Deviation. N = Sample size. Year = Year of publication of the study. k = number of reliability estimates.  $b_j$  = regression coefficient of each predictor. SE = Standard Error for  $b_j$ . LL and UL = lower and upper limits of the 95 % confidence interval for  $b_j$ . F = Knapp-Hartung's statistic for testing the significance of the predictor (the degrees of freedom for this statistic are 1 for the numerator and  $k - 2$  for the denominator). p = probability level for the F statistic.  $Q_E$  = statistic for testing the model misspecification.  $R^2$  = proportion of variance accounted for by the predictor. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

subscales.

#### 4. Discussion and conclusions

The purpose of this study was to conduct a reliability generalization meta-analysis on the CTQ-SF in order to provide a more comprehensive understanding of its reliability estimates. The main result of this meta-analysis is that the CTQ-SF exhibited acceptable reliability in the general (total score) scale and most of its subscales, but not all. The dimension or subscale that most commonly exhibited reliability problems was physical neglect, but there were also some suboptimal reliability indexes in emotional neglect and physical and emotional abuse. The variability exhibited in the reliability coefficients among studies was mainly associated to sample variability, with more variable samples exhibiting better estimates of reliability. Additionally, there were also some associations with geographical locations, financial status of the researcher, main researcher, and research design, but to a lesser extent. The CTQ-SF is one of the most used scales to measure and diagnose child maltreatment, and our results point out potential reliability problems in some dimensions and specific contexts in which care should be taken with this instrument.

Most of the studies only reported one type of reliability estimate which was, in most of the cases, Cronbach's alpha. However, there were also several studies reporting test-retest reliability. These results agree with prior RG meta-analyses (e.g., [Badenes-Ribera et al., 2023](#); [Rubio-Aparicio et al., 2020](#)). This outcome is clearly due to the long tradition of reporting this internal consistency coefficient although it has been demonstrated that other indexes perform better than alpha ([McNeish, 2018](#)).

The average reliability coefficient for CTQ-SF total scores, measured as Cronbach's alpha coefficient, was 0.891, and when assessed as McDonald's Omega coefficient, it was 0.800. These are considered as acceptable scores in both cases ([Cicchetti, 1994](#)). For the physical abuse, total scores were 0.843 and 894, respectively, for the emotional abuse total scores were 0.839 and 845, respectively, for the sexual abuse total scores were 0.916 and 900, for the physical neglect total scores were 0.656 and 0.740 respectively, and for emotional neglect total scores was 0.851 and 0.914 respectively. Following the psychometric theory ([Nunnally & Bernstein, 1994](#)), Cronbach's alpha estimates over 0.70 can be considered acceptable for exploratory research. Nevertheless, estimates higher than 0.80 are recommended for general research and higher than 0.90 in clinical practice. In addition, [Cicchetti \(1994\)](#) suggested the following guidelines to assess the clinical relevance of Cronbach's Alpha coefficients of test scores: unacceptable for coefficients lower than 0.7, fair for the range from 0.7 to 0.8, good for 0.8 to 0.9, and excellent for values over 0.9. Thus, according to these recommendations, the average internal consistency reliability measured as Cronbach's alpha coefficient of the CTQ-SF, physical abuse, emotional abuse, sexual abuse, and emotional neglect could be considered adequate for research and acceptable for making clinical decisions, when Cicchetti's guidelines are followed. However, the average internal consistency reliability for physical neglect can be considered unacceptable for exploratory and general research and clinical decisions. This result might be due to several factors: Firstly, there might be diversity in the perception of what is and what is not considered physical neglect. Individuals might interpret and recall physical neglect in diverse ways, leading to a lower consistency in responses. It is well known that neglect in general is one of the hardest types of maltreatment to assess ([Clément et al., 2016](#); [Hornor, 2014](#); [Turner et al., 2019](#)). Secondly, the concept of physical neglect may be culturally sensitive, leading to different results in different countries (and cultures) as cultural norms might affect the concept and interpretation of neglect, leading again to lower reliability across different cultural groups. In particular with physical neglect, the indicators in the subscale may be too orientated to the reality of Western countries, not only culturally but also economically ([Viola et al., 2016](#)). For example, items such as "The child/adolescent does not have enough to eat", "There is someone to take the child/adolescent to the doctor if he/she needs it" and "child/adolescent wears dirty clothes" may not be indicators of child maltreatment, but due to economic and cultural conditions. Therefore, in contexts (as in Africa and Asia) this would lower covariances with other indicators of the subscale and even produce a lack of unidimensionality, that turns into lower reliabilities. This explanation is congruent with the results of [Lansford et al. \(2015\)](#) who examined physical maltreatment and physical neglect among 13 different cultural groups in nine countries obtaining differences among cultures. Therefore, it can be concluded that with the specific construct of physical

**Table 4.**

Results of the multiple meta-regressions applied on Cronbach's alpha coefficients for the Emotional Abuse subscale total scores ( $k = 20$ ).

| Predictor variable                             | $b_j$                        | SE     | $t$    | $p$    |
|------------------------------------------------|------------------------------|--------|--------|--------|
| Intercept                                      | 7.559                        | 32.86  | 0.230  | 0.822  |
| Mean scores                                    | -0.048                       | 0.028  | -1.717 | 0.110  |
| Standard deviation of scores                   | 0.189                        | 0.083  | 2.290  | 0.039  |
| Sample size                                    | -0.0001                      | 0.0001 | -1.814 | 0.093  |
| Mean age                                       | 0.006                        | 0.013  | 0.508  | 0.620  |
| Standard deviation age                         | -0.005                       | 0.032  | -0.152 | 0.0882 |
| Publication year                               | -0.003                       | 0.016  | -0.186 | 0.855  |
| Global results:                                | $F(6, 13) = 4.369, p = .012$ |        |        |        |
| Total $\tau^2$ (intercept-only model): 0.112   | $R^2 = 0.540$                |        |        |        |
| Residual $\tau^2$ (full model): 0.034          | $Q_E(13) = 70.202, p < .001$ |        |        |        |
| Residual $\tau^2$ (one-predictor model): 0.051 |                              |        |        |        |

Note:  $b_j$  = regression coefficient of each predictor. SE = Standard Error for  $b_j$ .  $t$  = statistic for testing the significance of the predictor.  $p$  = probability level for the  $z$  statistic.  $F$  = Knapp-Hartung's statistic for testing the significance of the full model.  $Q_E$  = statistic for testing the model misspecification.  $R^2$  = proportion of variance accounted for by the predictors.

**Table 5.**

Results of the weighted ANOVAs applied on Cronbach's alpha coefficients for the CTQ-SF Scale total score and its five subscales, taking categorical moderator variables as independent variables.

| Variable                          | k  | $\alpha_+$ | 95 % CI |       | ANOVA results                   |
|-----------------------------------|----|------------|---------|-------|---------------------------------|
|                                   |    |            | LL      | LU    |                                 |
| <b>CTQ-SF TOTAL</b>               |    |            |         |       |                                 |
| <b>Research design</b>            |    |            |         |       |                                 |
| Cross-sectional                   | 15 | 0.896      | 0.873   | 0.915 | $F(1, 16) = 1.849, p = .193$    |
| Longitudinal                      | 3  | 0.856      | 0.770   | 0.910 | $R^2 = 0.043$                   |
| <b>Financial source</b>           |    |            |         |       |                                 |
| No/Not reported                   | 6  | 0.909      | 0.875   | 0.935 | $Q_w(16) = 678.220, p < .0001$  |
| Yes                               | 12 | 0.880      | 0.850   | 0.904 | $F(1, 16) = 2.262, p = .152$    |
| <b>Main researcher</b>            |    |            |         |       |                                 |
| Criminologist                     | 2  | 0.915      | 0.865   | 0.947 | $R^2 = 0.083$                   |
| Medical Psychological Center      | 4  | 0.830      | 0.766   | 0.876 | $Q_w(16) = 560.657, p < .0001$  |
| Other                             | 1  | 0.920      | 0.851   | 0.957 | $F(5, 12) = 3.017, p = .054$    |
| Psychiatrist                      | 3  | 0.875      | 0.812   | 0.917 | $R^2 = 0.407$                   |
| Psychologist                      | 6  | 0.909      | 0.880   | 0.931 | $Q_w(12) = 154.435, p < .0001$  |
| Social worker                     | 2  | 0.907      | 0.853   | 0.941 |                                 |
| <b>Location</b>                   |    |            |         |       |                                 |
| Asia                              | 7  | 0.848      | 0.808   | 0.880 | $F(2, 15) = 7.149, p = .007$    |
| Europe                            | 9  | 0.910      | 0.889   | 0.927 | $R^2 = 0.464$                   |
| North America                     | 2  | 0.915      | 0.868   | 0.945 | $Q_w(15) = 209.212, p < .0001$  |
| <b>Population Target</b>          |    |            |         |       |                                 |
| Clinical                          | 10 | 0.897      | 0.865   | 0.922 | $F(2, 15) = 0.458, p = .641$    |
| Community                         | 7  | 0.886      | 0.845   | 0.916 | $R^2 = 0.000$                   |
| Delinquents                       | 1  | 0.850      | 0.639   | 0.938 | $Q_w(15) = 613.359, p < .0001$  |
| <b>Clinical Population Target</b> |    |            |         |       |                                 |
| Clinical mental health            | 1  | 0.920      | 0.764   | 0.973 | $F(5, 4) = 1.123, p = .468$     |
| Depression                        | 3  | 0.864      | 0.741   | 0.929 | $R^2 = 0.073$                   |
| Personality disorders             | 1  | 0.940      | 0.817   | 0.980 | $Q_w(4) = 38.878, p < .0001$    |
| Psychiatric patients              | 1  | 0.920      | 0.758   | 0.974 |                                 |
| Schizophrenia                     | 3  | 0.874      | 0.753   | 0.936 |                                 |
| Substance abuse                   | 1  | 0.930      | 0.776   | 0.978 |                                 |
| <b>PHYSICAL ABUSE</b>             |    |            |         |       |                                 |
| <b>Research design</b>            |    |            |         |       |                                 |
| Cross-sectional                   | 31 | 0.842      | 0.814   | 0.866 | $F(1, 34) = 0.047, p = .830$    |
| Longitudinal                      | 5  | 0.849      | 0.773   | 0.900 | $R^2 = 0.000$                   |
| <b>Financial source</b>           |    |            |         |       |                                 |
| No/Not reported                   | 17 | 0.837      | 0.797   | 0.869 | $Q_w(34) = 1255.477, p < .0001$ |
| Yes                               | 19 | 0.849      | 0.814   | 0.878 | $F(1, 34) = 0.286, p = .596$    |
| <b>Main researcher</b>            |    |            |         |       |                                 |
| Criminologist                     | 2  | 0.860      | 0.732   | 0.927 | $R^2 = 0.000$                   |
| Medical Psychological Center      | 4  | 0.804      | 0.692   | 0.875 | $Q_w(34) = 1134.267, p < .0001$ |
| Other                             | 2  | 0.890      | 0.787   | 0.943 | $F(5, 30) = 0.664, p = .654$    |
| Psychiatrist                      | 9  | 0.860      | 0.809   | 0.897 | $R^2 = 0.000$                   |
| Psychologist                      | 17 | 0.837      | 0.797   | 0.870 | $Q_w(30) = 756.841, p < .0001$  |
| Social worker                     | 2  | 0.813      | 0.645   | 0.902 |                                 |
| <b>Location</b>                   |    |            |         |       |                                 |
| Africa                            | 2  | 0.780      | 0.578   | 0.885 | $F(4, 31) = 0.350, p = .842$    |
| Asia                              | 8  | 0.844      | 0.782   | 0.888 | $R^2 = 0.000$                   |
| Europe                            | 18 | 0.849      | 0.812   | 0.878 | $Q_w(31) = 818.708, p < .0001$  |
| North America                     | 7  | 0.840      | 0.774   | 0.887 |                                 |
| South America                     | 1  | 0.870      | 0.659   | 0.951 |                                 |
| <b>Study language</b>             |    |            |         |       |                                 |
| English                           | 35 | 0.842      | 0.816   | 0.865 | $F(1, 34) = 0.169, p = .683$    |
| Spanish                           | 1  | 0.870      | 0.668   | 0.949 | $R^2 = 0.000$                   |
| <b>Population Target</b>          |    |            |         |       |                                 |
| Clinical                          | 16 | 0.852      | 0.812   | 0.883 | $Q_w(34) = 1252.720, p < .0001$ |
| Community                         | 14 | 0.836      | 0.791   | 0.872 | $F(3, 32) = 0.142, p = .934$    |
| Delinquents                       | 3  | 0.832      | 0.708   | 0.903 | $R^2 = 0.000$                   |
| Mixed                             | 3  | 0.842      | 0.732   | 0.907 | $Q_w(32) = 1053.048, p < .0001$ |
| <b>Clinical Population Target</b> |    |            |         |       |                                 |
| Clinical mental health            | 3  | 0.873      | 0.840   | 0.900 | $F(5, 10) = 2.407, p = .111$    |
| Depression                        | 3  | 0.844      | 0.802   | 0.877 | $R^2 = 0.558$                   |
| Personality disorders             | 1  | 0.860      | 0.784   | 0.909 | $Q_w(10) = 25.878, p = .004$    |
| Psychiatric patients              | 3  | 0.878      | 0.851   | 0.900 |                                 |
| Schizophrenia                     | 3  | 0.825      | 0.768   | 0.867 |                                 |
| Substance abuse                   | 3  | 0.825      | 0.782   | 0.859 |                                 |
| <b>EMOTIONAL ABUSE</b>            |    |            |         |       |                                 |

(continued on next page)

Table 5. (continued)

| Variable                          | k  | $\alpha_+$ | 95 % CI |       | ANOVA results                   |
|-----------------------------------|----|------------|---------|-------|---------------------------------|
|                                   |    |            | LL      | LU    |                                 |
| <b>Research design</b>            |    |            |         |       | $F(1, 34) = 2.849, p = .101$    |
| Cross-sectional                   | 31 | 0.845      | 0.824   | 0.863 | $R^2 = 0.058$                   |
| Longitudinal                      | 5  | 0.795      | 0.720   | 0.850 | $Q_W(34) = 1258.549, p < .0001$ |
| <b>Financial source</b>           |    |            |         |       | $F(1, 34) = 10.288, p = .003$   |
| No/Not reported                   | 17 | 0.864      | 0.842   | 0.883 | $R^2 = 0.247$                   |
| Yes                               | 19 | 0.811      | 0.781   | 0.836 | $Q_W(34) = 640.771, p < .0001$  |
| <b>Main researcher</b>            |    |            |         |       | $F(5, 30) = 2.658, p = .042$    |
| Criminologist                     | 2  | 0.815      | 0.709   | 0.883 | $R^2 = 0.205$                   |
| Medical Psychological Center      | 4  | 0.735      | 0.640   | 0.806 | $Q_W(30) = 549.839, p < .0001$  |
| Other                             | 2  | 0.854      | 0.768   | 0.907 |                                 |
| Psychiatrist                      | 9  | 0.844      | 0.807   | 0.874 |                                 |
| Psychologist                      | 17 | 0.854      | 0.830   | 0.875 |                                 |
| Social worker                     | 2  | 0.846      | 0.760   | 0.901 |                                 |
| <b>Location</b>                   |    |            |         |       | $F(4, 31) = 5.128, p = .003$    |
| Africa                            | 2  | 0.718      | 0.576   | 0.812 | $R^2 = 0.361$                   |
| Asia                              | 8  | 0.784      | 0.734   | 0.826 | $Q_W(31) = 532.641, p < .0001$  |
| Europe                            | 18 | 0.858      | 0.837   | 0.876 |                                 |
| North America                     | 7  | 0.857      | 0.822   | 0.884 |                                 |
| South America                     | 1  | 0.850      | 0.715   | 0.921 |                                 |
| <b>Study language</b>             |    |            |         |       | $F(1, 34) = 0.039, p = .845$    |
| English                           | 35 | 0.838      | 0.818   | 0.857 | $R^2 = 0.000$                   |
| Spanish                           | 1  | 0.850      | 0.680   | 0.930 | $Q_W(34) = 1313.401, p < .0001$ |
| <b>Population Target</b>          |    |            |         |       | $F(3, 32) = 1.895, p = .150$    |
| Clinical                          | 16 | 0.859      | 0.832   | 0.882 | $R^2 = 0.073$                   |
| Community                         | 14 | 0.819      | 0.784   | 0.848 | $Q_W(32) = 879.078, p < .0001$  |
| Delinquents                       | 3  | 0.853      | 0.778   | 0.903 |                                 |
| Mixed                             | 3  | 0.802      | 0.711   | 0.865 |                                 |
| <b>Clinical Population Target</b> |    |            |         |       | $F(5, 10) = 1.205, p = .374$    |
| Clinical mental health            | 3  | 0.869      | 0.819   | 0.905 | $R^2 = 0.159$                   |
| Depression                        | 3  | 0.811      | 0.738   | 0.864 | $Q_W(10) = 42.845, p < .001$    |
| Personality disorders             | 1  | 0.870      | 0.770   | 0.927 |                                 |
| Psychiatric patients              | 3  | 0.873      | 0.829   | 0.906 |                                 |
| Schizophrenia                     | 3  | 0.846      | 0.779   | 0.892 |                                 |
| Substance abuse                   | 3  | 0.876      | 0.830   | 0.909 |                                 |
| <b>SEXUAL ABUSE</b>               |    |            |         |       |                                 |
| <b>Research design</b>            |    |            |         |       | $F(1, 35) = 0.914, p = .346$    |
| Cross-sectional                   | 31 | 0.919      | 0.899   | 0.935 | $R^2 = 0.000$                   |
| Longitudinal                      | 6  | 0.895      | 0.824   | 0.937 | $Q_W(35) = 3337.924, p < .0001$ |
| <b>Financial source</b>           |    |            |         |       | $F(1, 35) = 4.222, p = .047$    |
| No/Not reported                   | 17 | 0.932      | 0.909   | 0.949 | $R^2 = 0.091$                   |
| Yes                               | 20 | 0.899      | 0.868   | 0.922 | $Q_W(35) = 2199.070, p < .0001$ |
| <b>Main researcher</b>            |    |            |         |       | $F(5, 31) = 2.945, p = .027$    |
| Criminologist                     | 2  | 0.933      | 0.853   | 0.969 | $R^2 = 0.226$                   |
| Medical Psychological Center      | 4  | 0.797      | 0.648   | 0.882 | $Q_W(31) = 1085.713, p < .0001$ |
| Other                             | 3  | 0.885      | 0.782   | 0.940 |                                 |
| Psychiatrist                      | 9  | 0.918      | 0.881   | 0.943 |                                 |
| Psychologist                      | 17 | 0.930      | 0.909   | 0.947 |                                 |
| Social worker                     | 2  | 0.937      | 0.863   | 0.971 |                                 |
| <b>Location</b>                   |    |            |         |       | $F(4, 32) = 4.612, p = .005$    |
| Africa                            | 2  | 0.852      | 0.689   | 0.929 | $R^2 = 0.316$                   |
| Asia                              | 8  | 0.844      | 0.773   | 0.893 | $Q_W(32) = 927.568, p < .0001$  |
| Europe                            | 18 | 0.935      | 0.916   | 0.949 |                                 |
| North America                     | 8  | 0.926      | 0.893   | 0.949 |                                 |
| South America                     | 1  | 0.930      | 0.793   | 0.976 |                                 |
| <b>Study language</b>             |    |            |         |       | $F(1, 35) = 0.089, p = .767$    |
| English                           | 36 | 0.915      | 0.895   | 0.931 | $R^2 = 0.000$                   |
| Spanish                           | 1  | 0.930      | 0.747   | 0.981 | $Q_W(35) = 3334.660, p < .0001$ |
| <b>Population Target</b>          |    |            |         |       | $F(3, 35) = 0.574, p = .636$    |
| Clinical                          | 17 | 0.913      | 0.881   | 0.936 | $R^2 = 0.075$                   |
| Community                         | 14 | 0.909      | 0.873   | 0.935 | $Q_W(33) = 2737.990, p < .0001$ |
| Delinquents                       | 3  | 0.945      | 0.885   | 0.974 |                                 |
| Mixed                             | 3  | 0.924      | 0.843   | 0.963 |                                 |
| <b>Clinical Population Target</b> |    |            |         |       | $F(5, 11) = 1.762, p = .201$    |
| Clinical mental health            | 4  | 0.916      | 0.843   | 0.955 | $R^2 = 0.191$                   |
| Depression                        | 3  | 0.804      | 0.593   | 0.906 | $Q_W(11) = 324.354, p < .0001$  |
| Personality disorders             | 1  | 0.910      | 0.682   | 0.975 |                                 |
| Psychiatric patients              | 3  | 0.937      | 0.870   | 0.969 |                                 |
| Schizophrenia                     | 3  | 0.910      | 0.811   | 0.958 |                                 |

(continued on next page)

Table 5. (continued)

| Variable                          | k  | $\alpha_+$ | 95 % CI |       | ANOVA results                   |
|-----------------------------------|----|------------|---------|-------|---------------------------------|
|                                   |    |            | LL      | LU    |                                 |
| Substance abuse                   | 3  | 0.944      | 0.884   | 0.973 |                                 |
| <b>PHYSICAL NEGLECT</b>           |    |            |         |       |                                 |
| <b>Research design</b>            |    |            |         |       | $F(1, 34) = 0.080, p = .783$    |
| Cross-sectional                   | 31 | 0.658      | 0.604   | 0.704 | $R^2 = 0.000$                   |
| Longitudinal                      | 5  | 0.639      | 0.480   | 0.750 | $Q_W(34) = 749.524, p < .0001$  |
| <b>Financial source</b>           |    |            |         |       | $F(1, 34) = 0.004, p = .953$    |
| No/Not reported                   | 17 | 0.654      | 0.579   | 0.715 | $R^2 = 0.000$                   |
| Yes                               | 19 | 0.656      | 0.586   | 0.715 | $Q_W(34) = 713.358, p < .0001$  |
| <b>Main researcher</b>            |    |            |         |       | $F(5, 30) = 1.816, p = .140$    |
| Criminologist                     | 2  | 0.810      | 0.676   | 0.889 | $R^2 = 0.133$                   |
| Medical Psychological Center      | 4  | 0.530      | 0.322   | 0.675 | $Q_W(30) = 569.459, p < .0001$  |
| Other                             | 2  | 0.688      | 0.465   | 0.819 |                                 |
| Psychiatrist                      | 9  | 0.638      | 0.534   | 0.719 |                                 |
| Psychologist                      | 17 | 0.669      | 0.602   | 0.724 |                                 |
| Social worker                     | 2  | 0.590      | 0.307   | 0.757 |                                 |
| <b>Location</b>                   |    |            |         |       | $F(4, 31) = 1.286, p = .297$    |
| Africa                            | 2  | 0.607      | 0.320   | 0.773 | $R^2 = 0.042$                   |
| Asia                              | 8  | 0.633      | 0.513   | 0.723 | $Q_W(31) = 610.687, p < .0001$  |
| Europe                            | 18 | 0.644      | 0.572   | 0.704 |                                 |
| North America                     | 7  | 0.729      | 0.638   | 0.798 |                                 |
| South America                     | 1  | 0.410      | -0.369  | 0.746 |                                 |
| <b>Study language</b>             |    |            |         |       | $F(1, 34) = 1.716, p = .199$    |
| English                           | 35 | 0.660      | 0.611   | 0.702 | $R^2 = 0.019$                   |
| Spanish                           | 1  | 0.410      | -0.371  | 0.746 | $Q_W(34) = 793.430, p < .0001$  |
| <b>Population Target</b>          |    |            |         |       | $F(3, 32) = 0.232, p = .874$    |
| Clinical                          | 16 | 0.673      | 0.597   | 0.735 | $R^2 = 0.000$                   |
| Community                         | 14 | 0.637      | 0.550   | 0.708 | $Q_W(32) = 716.859, p < .0001$  |
| Delinquents                       | 3  | 0.622      | 0.380   | 0.769 |                                 |
| Mixed                             | 3  | 0.671      | 0.476   | 0.794 |                                 |
| <b>Clinical Population Target</b> |    |            |         |       | $F(5, 10) = 1.821, p = .196$    |
| Clinical mental health            | 3  | 0.616      | 0.445   | 0.734 | $R^2 = 0.312$                   |
| Depression                        | 3  | 0.550      | 0.349   | 0.689 | $Q_W(10) = 60.670, p < .0001$   |
| Personality disorders             | 1  | 0.810      | 0.638   | 0.900 |                                 |
| Psychiatric patients              | 3  | 0.700      | 0.577   | 0.787 |                                 |
| Schizophrenia                     | 3  | 0.705      | 0.559   | 0.803 |                                 |
| Substance abuse                   | 3  | 0.700      | 0.570   | 0.790 |                                 |
| <b>EMOTIONAL NEGLECT</b>          |    |            |         |       |                                 |
| <b>Research design</b>            |    |            |         |       | $F(1, 33) = 5.695, p = .023$    |
| Cross-sectional                   | 30 | 0.861      | 0.838   | 0.881 | $R^2 = 0.120$                   |
| Longitudinal                      | 5  | 0.774      | 0.669   | 0.846 | $Q_W(33) = 1138.346, p < .0001$ |
| <b>Financial source</b>           |    |            |         |       | $F(1, 33) = 24.442, p < .001$   |
| No/Not reported                   | 16 | 0.891      | 0.870   | 0.908 | $R^2 = 0.442$                   |
| Yes                               | 19 | 0.806      | 0.772   | 0.835 | $Q_W(33) = 549.850, p < .0001$  |
| <b>Main researcher</b>            |    |            |         |       | $F(5, 29) = 4.394, p = .004$    |
| Criminologist                     | 2  | 0.655      | 0.417   | 0.796 | $R^2 = 0.351$                   |
| Medical Psychological Center      | 4  | 0.801      | 0.715   | 0.861 | $Q_W(29) = 622.976, p < .0001$  |
| Other                             | 2  | 0.869      | 0.778   | 0.923 |                                 |
| Psychiatrist                      | 9  | 0.828      | 0.779   | 0.866 |                                 |
| Psychologist                      | 16 | 0.880      | 0.856   | 0.901 |                                 |
| Social worker                     | 2  | 0.886      | 0.809   | 0.932 |                                 |
| <b>Location</b>                   |    |            |         |       | $F(4, 30) = 1.750, p = .165$    |
| Africa                            | 2  | 0.778      | 0.592   | 0.880 | $R^2 = 0.080$                   |
| Asia                              | 8  | 0.818      | 0.751   | 0.866 | $Q_W(30) = 786.119, p < .0001$  |
| Europe                            | 17 | 0.875      | 0.845   | 0.899 |                                 |
| North America                     | 7  | 0.847      | 0.788   | 0.889 |                                 |
| South America                     | 1  | 0.790      | 0.478   | 0.915 |                                 |
| <b>Study language</b>             |    |            |         |       | $F(1, 33) = 0.566, p = .457$    |
| English                           | 34 | 0.853      | 0.828   | 0.874 | $R^2 = 0.000$                   |
| Spanish                           | 1  | 0.790      | 0.459   | 0.918 | $Q_W(33) = 1201.911, p < .0001$ |
| <b>Population Target</b>          |    |            |         |       | $F(3, 31) = 1.854, p = .157$    |
| Clinical                          | 16 | 0.863      | 0.829   | 0.890 | $R^2 = 0.079$                   |
| Community                         | 13 | 0.818      | 0.769   | 0.856 | $Q_W(31) = 902.097, p < .0001$  |
| Delinquents                       | 3  | 0.883      | 0.802   | 0.930 |                                 |
| Mixed                             | 3  | 0.886      | 0.812   | 0.930 |                                 |
| <b>Clinical Population Target</b> |    |            |         |       | $F(5, 10) = 0.618, p = .690$    |
| Clinical mental health            | 3  | 0.851      | 0.763   | 0.906 | $R^2 = 0.000$                   |
| Depression                        | 3  | 0.834      | 0.735   | 0.896 | $Q_W(10) = 69.205, p < .0001$   |
| Personality disorders             | 1  | 0.900      | 0.776   | 0.955 |                                 |

(continued on next page)

Table 5. (continued)

| Variable             | k | $\alpha_+$ | 95 % CI |       | ANOVA results |
|----------------------|---|------------|---------|-------|---------------|
|                      |   |            | LL      | LU    |               |
| Psychiatric patients | 3 | 0.885      | 0.821   | 0.927 |               |
| Schizophrenia        | 3 | 0.846      | 0.748   | 0.906 |               |
| Substance abuse      | 3 | 0.876      | 0.805   | 0.922 |               |

Note. ANOVA = Analysis of variance.  $k$  = number of reliability estimates.  $\alpha_+$  = mean coefficient alpha. LL and LU = lower and upper 95 % confidence limits for  $\alpha_+$ .  $F$  = 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.

neglect, it can be harder to establish a clear and objective definition. Thus, physical neglect scores should be interpreted with caution when this instrument is used in both research and clinical contexts.

Concerning to test-retest reliability estimates, Pearson's correlation coefficients obtained an average of 0.788 for the CTQ-SF total scores, 0.709 for the physical abuse total scores, 0.704 for the emotional abuse total scores, 0.677 for the sexual abuse total scores, 0.668 for physical neglect total scores, and 0.681 for the emotional neglect total scores. For interpreting the adequacy of test-retest estimates it must be considered the time interval between test and retest administration (Charter, 2003; Revelle & Condon, 2018; Watson, 2004). Chmielewski and Watson (2009) pointed out that a time lag of two weeks is an index of dependability and not an index of temporal stability. However, Revelle and Condon (2018) pointed out that a time interval of 1–4 weeks between tests is indicative of temporal stability and a short term of 45 min is indicative of dependability. In this RG meta-analysis, the studies used time intervals ranging from 2 to 141 weeks (only one study used 2 weeks), according to the guidelines of Chmielewski and Watson (2009), the values of test-retest correlations might be considered an index of temporal stability. In addition, given that the average test-retest reliabilities ranged from 0.688 to 0.788, these can be considered as indicative of temporal stability (Revelle & Condon, 2018; Watson, 2004). Thus, all measures showed adequate stability across time.

In addition, the reliability estimates presented large heterogeneity across studies that might be not explained by sampling error alone. This is indicative that reliability of test scores changes depending on the context where the CTQ-SF or its five subscales were applied, the study characteristics, and the composition of the samples (target population, variability, etc.). This result might be due to the fact that maybe translating an instrument to another language does not necessarily mean that it will be adapted to the new culture. There is considerable variation across cultures about what constitutes child maltreatment. For example, a survey of 72 countries (Daro, 2006) found great geographical variability in recognizing behaviors such as physical discipline as abusive. In short, there are no universal parenting norms or definitions of maltreatment (Korbin, 1991). Differences in translation quality or cultural relevance may affect the reliability of the instrument (Bradley, 2013). Additionally, diversity in the composition of study samples, such as age groups, socio-economic status, or belonging or not to clinical populations, can also influence the reliability of the CTQ-SF. Different subgroups may interpret and respond to the questionnaire differently. The same could occur with differences in study designs as data collection methods, administration procedures and changes over time (as events from the past are being recalled) may also contribute to heterogeneity. Consequently, we encourage researchers to estimate the reliability of their sample every time that they use CTQ-SF or its five subscales.

Moderator analyses revealed that the variability of total scores in CTQ-SF, physical abuse, emotional abuse, sexual abuse, and physical neglect exhibited a statistically significant association with Cronbach's alpha coefficients. Thus, the Cronbach's Alpha coefficients obtained in the studies were higher when the variability of scores observed was larger. This is in line with the well-known criticisms on Cronbach's alpha as a measure of internal consistency (Yang & Green, 2011). Alpha is a biased estimate unless items are unidimensional and tau-equivalent, assumptions seldom found in practice (Jöreskog, 1971), and definitively not a feature of the CTQ-SF (Tomás et al., 2023). Additionally, Cronbach's alpha builds on the assumption that the responses to individual questions are normally distributed, and have equal variance, but more importantly, a well-known property of alpha is that depends on the variability of what is measured in the population (sample) analyzed so that the more variability, the higher the reliability (Viladrich et al., 2017). This is exactly the pattern of results we have found in this meta-analysis. On the contrary, omega is more robust than alpha against deviations from these assumptions and therefore a more suitable measure of internal consistency (Gabler & Raykov, 2017).

For emotional abuse subscale, in addition to the variability of the total scores, the participants' mean age, the variability of participants' age, and the year of publication of the study showed a statistically significant association with Cronbach's alpha coefficients. A multiple meta-regression analysis showed only the variability of scores showed a statistically significant association with Cronbach's alpha coefficients after controlling for the influence of the other variables, with 53.8 % of the variance explained. Thus, the Cronbach's Alpha coefficients obtained in the studies were higher when the variability of scores observed was larger. In addition, the geographical location, that is, where the study was conducted exhibited a relationship with the Cronbach's Alpha coefficients in the CTQ-SF total score, emotional abuse, and sexual abuse. Overall, the average reliability was lower for studies conducted in Africa and Asia. Finally, the financial source and the affiliation of the main researcher of the study showed a relationship with Cronbach's Alpha coefficients in the emotional abuse, sexual abuse, and emotional neglect subscale. The average reliability was lower for studies that were financed

than those that were not financed and for studies conducted by people who belonged to the Medical Psychological Center, compared to studies conducted by other professionals. In light of these results, the positive relationship between alpha coefficients and the variability of scores suggests that when there is more diversity in responses, the internal consistency tends to be higher. Therefore, these results suggest that the CTQ-SF might be more reliable in capturing nuances when there is a broader range of experiences among respondents. Additionally, the country differences in reliability, as in previous results, clearly indicate potential cultural and contextual influences. Differences in translation, adaptation, or cultural norms might play a role, but also, this result is to be expected taking into account that the original CTQ-SF was developed in the United States, as the instrument might more accurately reflect North American society than African and Asian societies.

Additionally, the relationship between financial sources and reliability in certain subscales may be attributed to budgetary and time constraints that affect the quality of the research process. Finally, results regarding the main researcher's affiliation, variations in the expertise, and familiarity with the instrument and/or cultural context of the sample may impact the quality of data collection and interpretation.

## 5. Limitations and directions for future research

One of the main limitations of this study was focusing only on psychometric validations studies of the CTQ-SF. Consequently, future RG meta-analyses should consider any empirical studies that have used some version of the CTQ-SF and provided reliability estimates. The second limitation was the small number of studies that provided test-retest reliabilities as a measure of temporal stability of test scores and McDonald's Omega estimates as a measure of internal consistency, which limited the generalizability of RG metanalytic results on these types of reliability of the test scores. In addition, this fact (small number of studies), did not allow us to analyze potential moderator variables related to these two types of reliability measures. Finally, we could not examine the reliability of the CTQ-SF in professional settings as there is little to no education for professionals in child welfare in this instrument, showing the need to provide better education for professionals in maltreatment instruments (Tarín-Cayuela, 2022).

Future research should explore refining or proposing alternative measures for the physical neglect scale that might present better reliability across samples. In this line, combining these quantitative findings with qualitative information that attempts to explain these differences can provide a deeper understanding of respondents' perceptions of physical neglect.

In conclusion, this study contributes to the understanding of reliability estimates in a widely used child maltreatment instrument such as the CTQ-SF. Our results mainly support the reliability of this instrument across studies with a few comments: Firstly, findings related to physical neglect should be further explored and interpreted with caution; and secondly, researchers should consider cultural differences in the interpretation of child maltreatment experiences in order to adapt not only the language but also the meaning to the sample in which is being used.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.chiabu.2024.106941>.

## Funding

This work was supported by the University of Valencia (OTR2021-07233CATED). Additionally, the researcher Sylvia Georgieva is receiving a PhD grant from the Valencian Government (ACIF/2021/175).

## CRediT authorship contribution statement

**Laura Badenes-Ribera:** Writing – original draft, Formal analysis, Conceptualization. **Sylvia Georgieva:** Writing – original draft, Investigation, Data curation. **Jose M. Tomás:** Writing – review & editing, Supervision, Investigation, Funding acquisition. **José Javier Navarro-Pérez:** Writing – review & editing, Visualization, Supervision, Project administration.

## Declaration of competing interest

None.

## Data availability

Data is shared in a public web which link is embeded in the text

## Appendix A

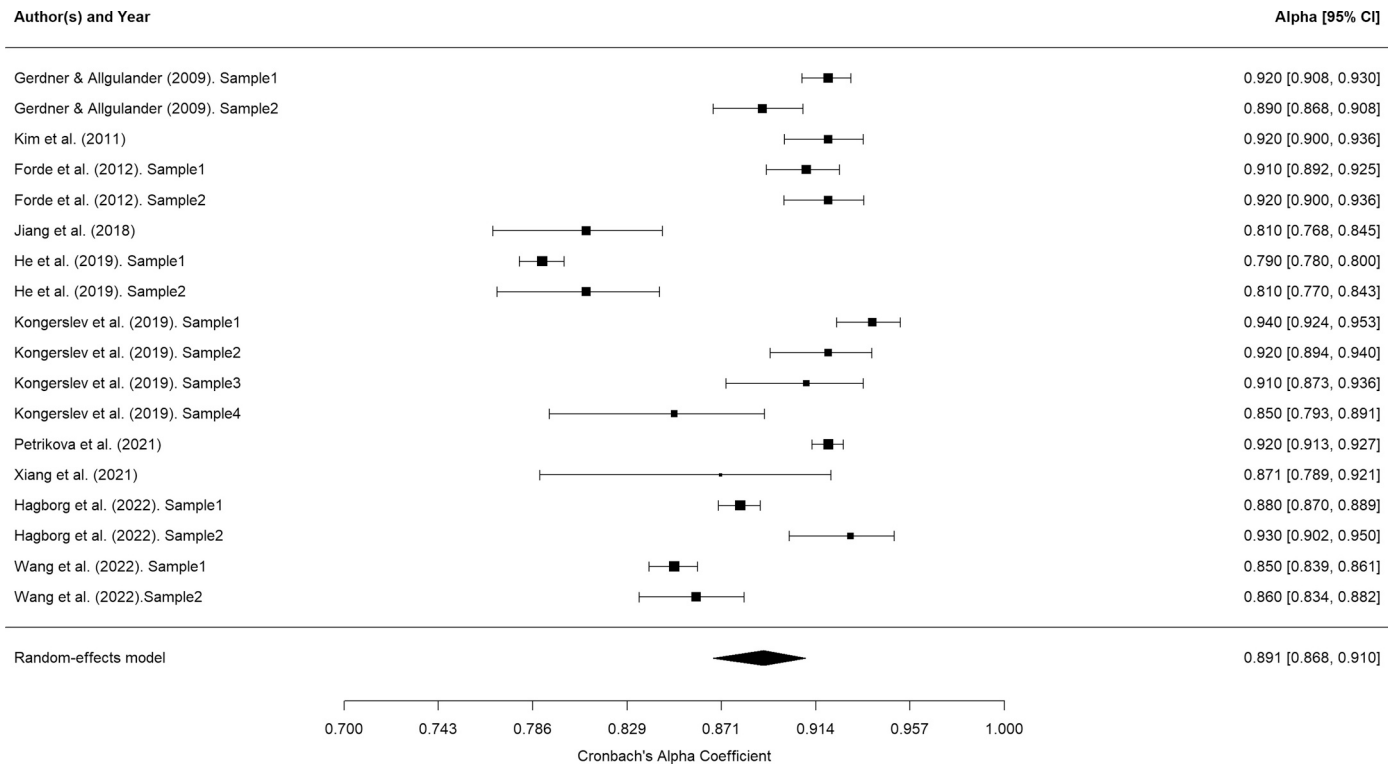


Fig. 2. Forest plot displaying Cronbach's alpha coefficients (and their 95 % confidence intervals) for the CTQ-SF total score.

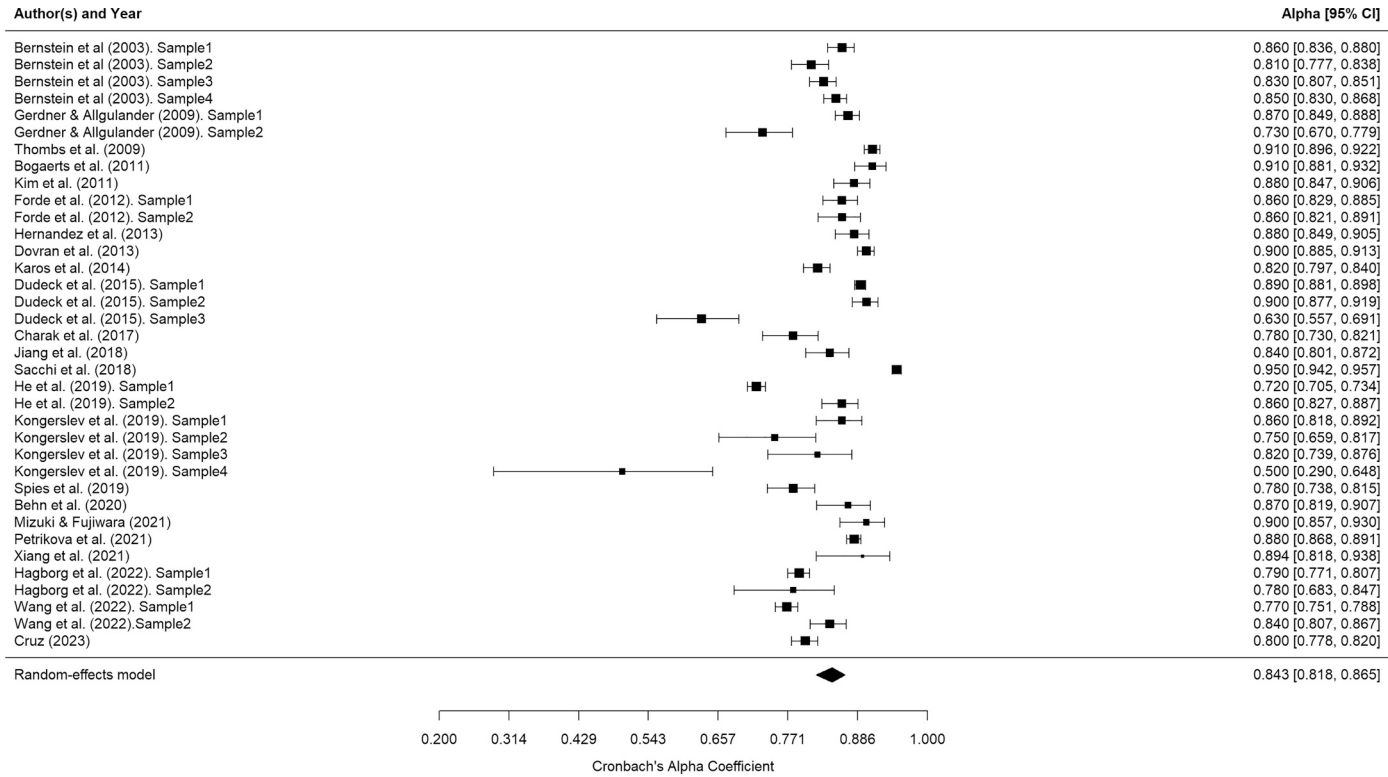
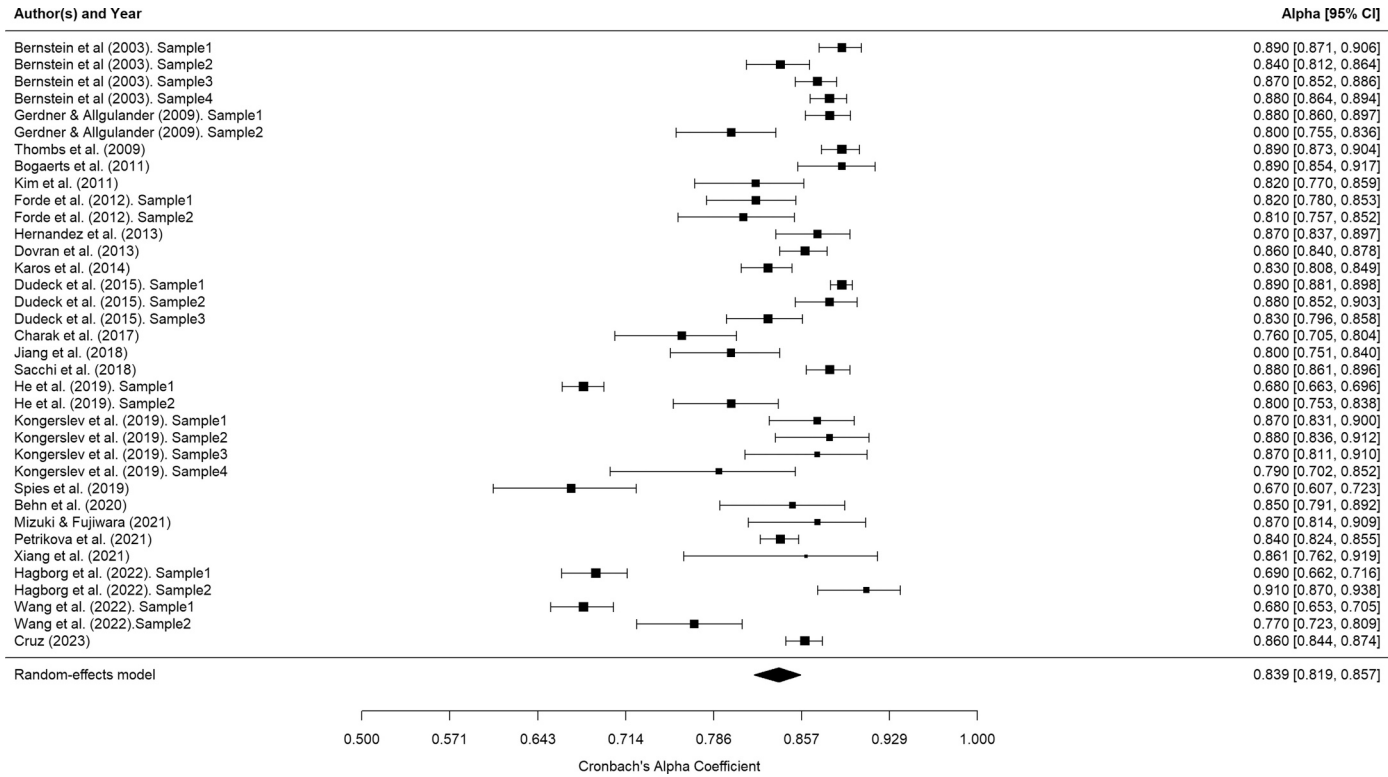
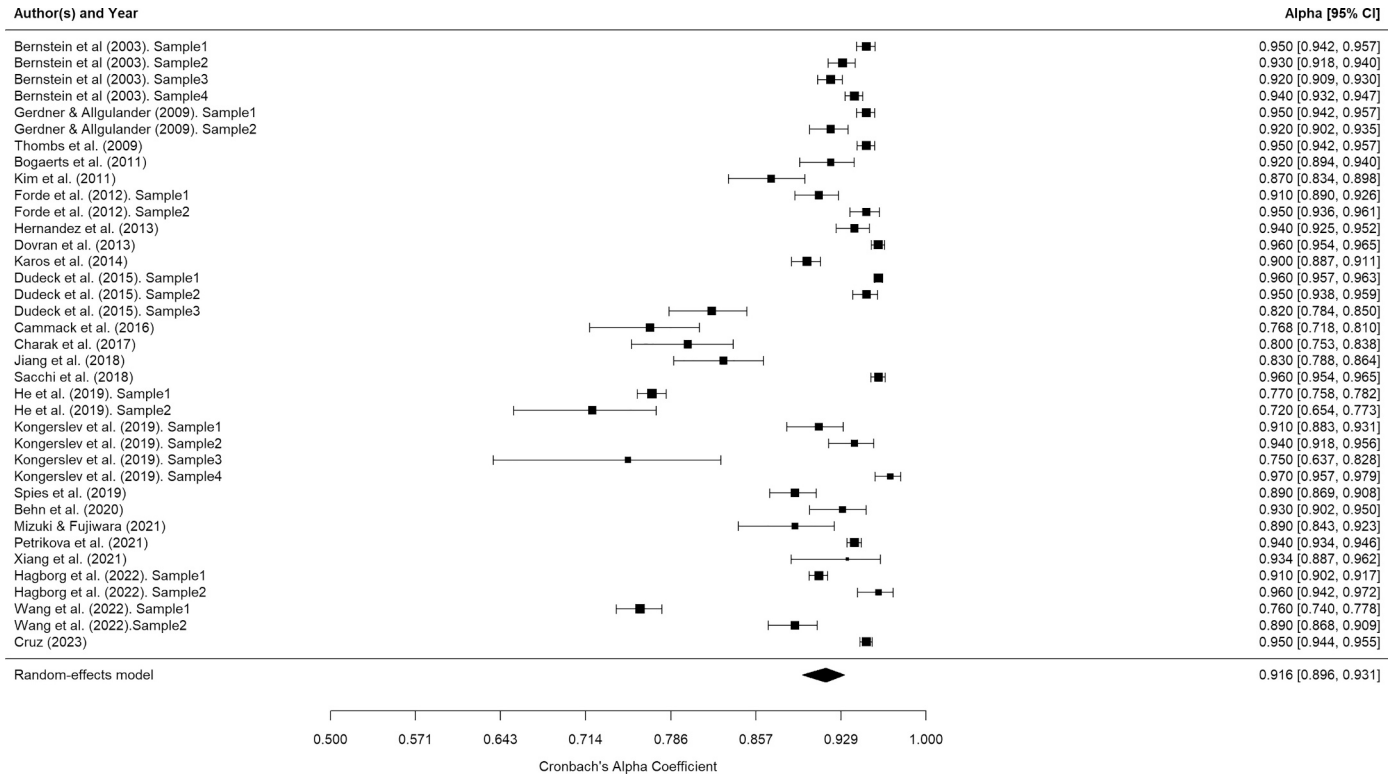


Fig. 3. Forest plot displaying Cronbach's alpha coefficients (and their 95 % confidence intervals) for the Physical abuse total score.



**Fig. 4.** Forest plot displaying Cronbach's alpha coefficients (and their 95 % confidence intervals) for the Emotional abuse total score.



**Fig. 5.** Forest plot displaying Cronbach's alpha coefficients (and their 95 % confidence intervals) for the Sexual abuse total score.

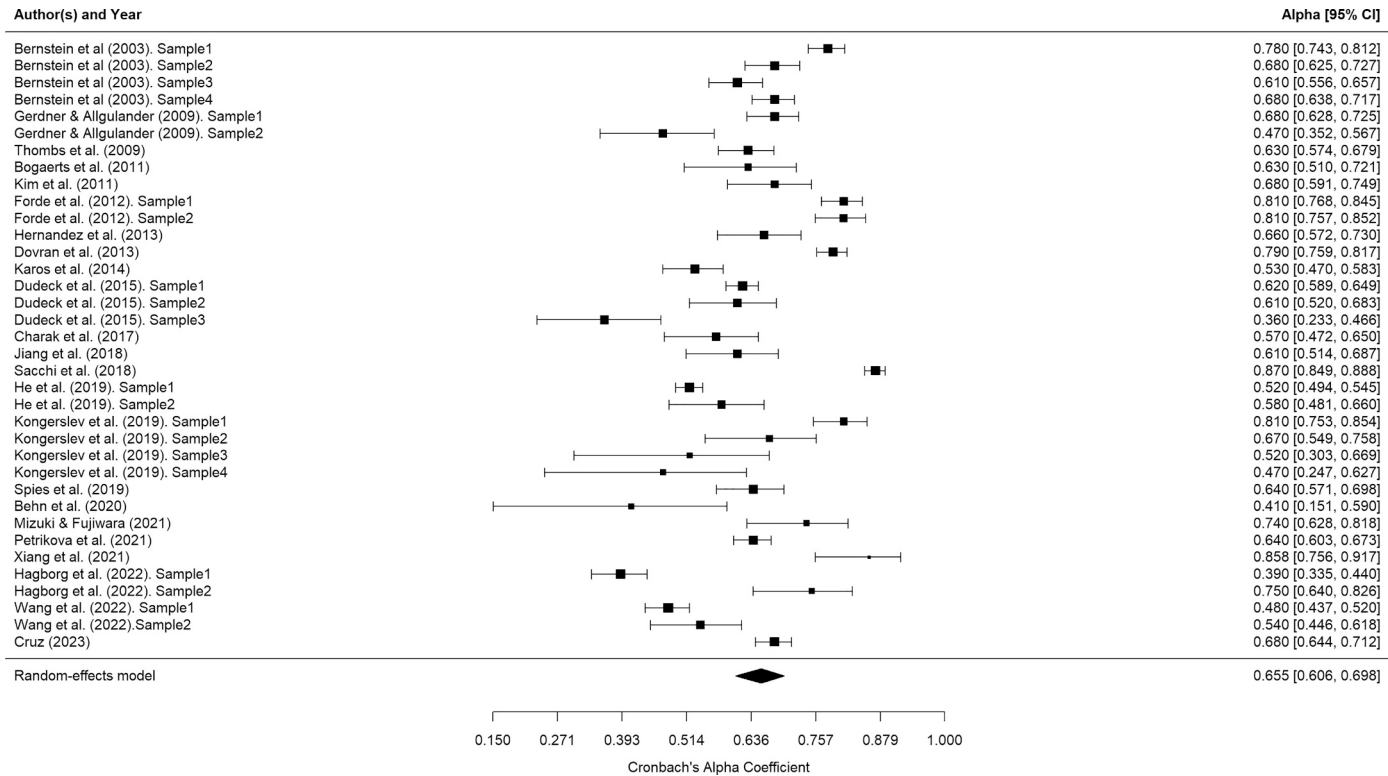


Fig. 6. Forest plot displaying Cronbach's alpha coefficients (and their 95 % confidence intervals) for the Physical neglect total score.

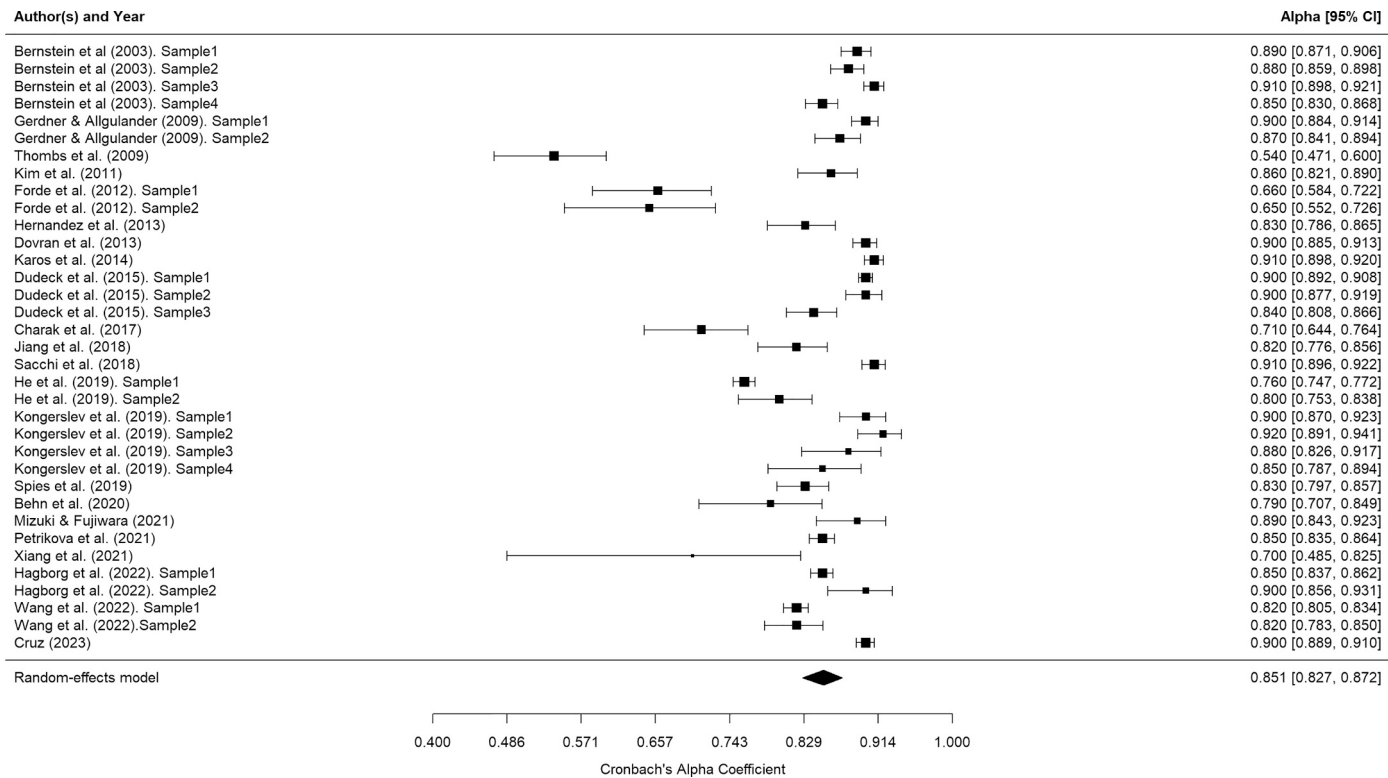


Fig. 7. Forest plot displaying Cronbach's alpha coefficients (and their 95 % confidence intervals) for the Emotional neglect total score.

References<sup>1</sup>

- Badenes-Ribera, L., Carmen Duro-García, C., López-Ibáñez, C., Martí-Vilar, M., & Sánchez-Meca. (2023). The adult prosocialness behavior scale: A reliability generalization meta-analysis. *International Journal of Behavioral Development*, 47(1), 59–71. <https://doi.org/10.1177/01650254221128280>
- Baker, A. J. L., & Maiorino, E. (2010). Assessments of emotional abuse and neglect with the CTQ: Issues and estimates. *Children and Youth Services Review*, 32, 740–748. <https://doi.org/10.1016/j.childyouth.2010.01.011>
- \*Behn, A., Vöhringer, P. A., Martínez, P., Domínguez, A. P., González, A., Carrasco, M. I., & Gloger, S. (2020). Validación de la versión en español del Childhood Trauma Questionnaire-Short Form en Chile, en una muestra de pacientes con depresión clínica. *Revista Médica de Chile*, 148(3), 336–343. <https://doi.org/10.4067/S0034-98872020000300336>
- Bekhet, A. K., & Zauszniewski, J. A. (2012). Methodological triangulation: An approach to understanding data. *Nurse Researcher*, 20(2), 40–43. <https://doi.org/10.7748/nr2012.11.20.2.40.c9442>
- Bernstein, D. P., & Fink, L. (1998). *Childhood trauma questionnaire: A retrospective self-report manual*. The Psychological Corporation.
- \*Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., ... Zule, W. (2003). Development and validation of a brief screening version of the childhood trauma questionnaire. *Child Maltreatment & Neglect*, 27(2), 169–190. [https://doi.org/10.1016/S0145-2134\(02\)00541-0](https://doi.org/10.1016/S0145-2134(02)00541-0)
- 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>
- Borenstein, M. J., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.
- Bradley, C. L. A. R. E. (2013). Translation of questionnaires for use in different languages and cultures. In *Handbook of psychology and diabetes* (pp. 43–55). Routledge. <https://doi.org/10.4324/9781315077369>
- 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>
- Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, 97(1), 186–202. <https://doi.org/10.1037/a0015618>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessments instruments in psychology. *Psychological Assessment*, 6, 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- Clément, M.É., Bérubé, A., & Chamberland, C. (2016). Prevalence and risk factors of child neglect in the general population. *Public Health*, 138, 86–92. <https://doi.org/10.1016/j.puhe.2016.03.018>
- Daro, D. (2006). *World perspectives on child abuse* (7th ed.). West Chicago, IL: ISPCAN.
- \*Dovran, A., Winje, D., Øverland, S. N., Breivik, K., Arefjord, K., Dalsbo, A. S., ... Waage, L. (2013). Psychometric properties of the Norwegian version of the Childhood Trauma Questionnaire in high-risk groups. *Scandinavian Journal of Psychology*, 54(4), 286–291. <https://doi.org/10.1111/sjop.12052>
- \*Forde, D. R., Baron, S. W., Scher, C. D., & Stein, M. B. (2012). Factor structure and reliability of the childhood trauma questionnaire and prevalence estimates of trauma for male and female street youth. *Journal of Interpersonal Violence*, 27(2), 364–379. <https://doi.org/10.1177/0886260511416461>
- Gabler, S., & Raykov, T. (2017). Evaluation of maximal reliability for unidimensional measuring instruments with correlated errors. *Structural Equation Modeling*, 24, 104–111. <https://doi.org/10.1080/10705511.2016.1159916>
- Georgieva, S., Tomas, J. M., & Navarro-Pérez, J. J. (2021). Systematic review and critical appraisal of childhood trauma questionnaire-short form (CTQ-SF). *Child Abuse & Neglect*, 120, Article 105223. <https://doi.org/10.1016/j.chiabu.2021.105223>
- Georgieva, S., Tomás, J. M., Navarro-Pérez, J. J., & Samper-García, P. (2022). Systematic review and critical appraisal of five of the most recurrently validated child maltreatment assessment instruments from 2010 to 2020. *Trauma, Violence & Abuse*, 0, 1–18. <https://doi.org/10.1177/15248380221097694>
- 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>
- 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>
- Hornor, G. (2014). Child neglect: Assessment and intervention. *Journal of Pediatric Health Care*, 28(2), 186–192. <https://doi.org/10.1016/j.pedhc.2013.10.002>
- Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I<sup>2</sup> index? *Psychological Methods*, 11, 193–206. <https://doi.org/10.1037/1082-989X.11.2.193>
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109–133.
- \*Kongerslev, M. T., Bach, B., Rossi, G., Trauelsen, A. M., Ladegaard, N., Løkkegaard, S. S., & Bo, S. (2019). Psychometric validation of the Childhood Trauma Questionnaire-Short Form (CTQ-SF) in a Danish clinical sample. *Child Abuse & Neglect*, 94, 104026. <https://doi.org/https://doi.org/10.1016/j.chiabu.2019.104026>
- Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in medicine*, 22(17), 2693–2710.
- Korbin, J. E. (1991). Cross-cultural perspectives and research directions for the 21st century. *Child Abuse & Neglect*, 15(Suppl. 1), 67–77.
- Lansford, J. E., Godwin, J., Tirado, L. M. U., Zelli, A., Al-Hassan, S. M., Bacchini, D., ... Alampay, L. P. (2015). Individual, family, and culture level contributions to child physical abuse and neglect: A longitudinal study in nine countries. *Development and Psychopathology*, 27(4pt2), 1417–1428. <https://doi.org/10.1017/S095457941500084X>
- 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>
- Madison, R. S., Narayan, A. J., Atzl, V. M., Rivera, L. M., & Lieberman, A. F. (2020). Childhood maltreatment on the adverse childhood experiences (ACEs) scale versus the childhood trauma questionnaire (CTQ) in a perinatal sample. *Journal of Aggression, Maltreatment & Trauma*, 29(1), 38–56. <https://doi.org/10.1080/10926771.2018.1524806>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1), 1–9. <https://doi.org/10.1186/2046-4053-4-1>
- Mokkink, L. B., De Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, 27(5), 1171–1179. <https://doi.org/10.1007/s11136-017-1765-4>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw Hill.
- Prinsen, C. A., Mokkink, L. B., Bouter, L. M., Alonso, J., Patrick, D. L., De Vet, H. C., & Terwee, C. B. (2018). COSMIN guideline for systematic reviews of patient-reported outcome measures. *Quality of Life Research*, 27(5), 1147–1157. <https://doi.org/10.1007/s11136-018-1798-3>
- 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.
- Revelle, W., & Condon, D. (2018). Reliability from  $\alpha$  to  $\omega$ : A tutorial. Retrieved from <https://psyarxiv.com/2y3w9>.
- 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>

<sup>1</sup> Studies included in the RG meta-analysis have been marked with an asterisk (\*).

- Saini, S. M., Hoffmann, C. R., Pantelis, C., Everall, I. P., & Bousman, C. A. (2019). Systematic review and critical appraisal of child maltreatment measurement instruments. *Psychiatry Research*, 272, 106–113. <https://doi.org/10.1016/j.psychres.2018.12.068>
- Sánchez-Meca, J., López-López, J. A., & López-Pina, J. A. (2013). Some recommended statistical analytic practices when reliability generalization studies are conducted. *The British Journal of Mathematical and Statistical Psychology*, 66(3), 402–425. <https://doi.org/10.1111/j.2044-8317.2012.02057.x>
- Sánchez-Meca, J., & 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-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>
- Schmidt, M. R., Narayan, A. J., Atzl, V. M., Rivera, L. M., & Lieberman, A. F. (2020). Childhood maltreatment on the Adverse Childhood Experiences (ACEs) Scale versus the Childhood Trauma Questionnaire (CTQ) in a perinatal sample. *Journal of Aggression, Maltreatment & Trauma*, 29(1), 38–56. <https://doi.org/10.1080/10926771.2018.1524806>
- Tarín-Cayuela, M. (2022). Las necesidades de formación de las educadoras y los educadores sociales en el ámbito de la infancia y la adolescencia vulnerable. *Research in Education and Learning Innovation Archives*, 29, 13–26.
- Teicher, M. H., & Parigger, A. (2015). The “Maltreatment and Maltreatment Chronology of Exposure”(MACE) scale for the retrospective assessment of maltreatment and neglect during development. *PLoS One*, 10(2), Article e0117423. <https://doi.org/10.1371/journal.pone.0117423>
- Terwee, C. B., Prinsen, C. A., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., ... Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. *Quality of Life Research*, 27(5), 1159–1170. <https://doi.org/10.1007/s11136-018-1829-0>
- Tomás, J. M., Fernández, I., Navarro-Pérez, J., & Carbonell, A. (2023). Adaptation of the Spanish-version of the childhood trauma questionnaire-short form (CTQ-SF) as an informant-based instrument. *Journal of Social Service Research*. <https://doi.org/10.1080/01488376.2023.2283578>
- Turner, H. A., Vanderminde, J., Finkelhor, D., & Hamby, S. (2019). Child neglect and the broader context of child victimization. *Child Maltreatment*, 24(3), 265–274. <https://doi.org/10.1177/1077559518825312>
- 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>
- Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología*, 33, 755–782. <https://doi.org/10.6018/analesps.33.3.268401>
- Viola, T. W., Salum, G. A., Kluwe-Schiavon, B., Sanvicente-Vieira, B., Levandowski, M. L., & Grassi-Oliveira, R. (2016). The influence of geographical and economic factors in estimates of childhood abuse and neglect using the Childhood Trauma Questionnaire: A worldwide meta-regression analysis. *Child Abuse & Neglect*, 51, 1–11. <https://doi.org/10.1016/j.chiabu.2015.11.019>
- 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>
- WHO. (2020). Child maltreatment. <https://www.who.int/news-room/fact-sheets/detail/child-maltreatment>.
- Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, 29, 377–392.